# Auditoría de Proyectos Big Data, Cloud Computing y Open Data


**TRABAJO FIN DE GRADO EN INGENIERÍA INFORMÁTICA**

**Berta Montes Cobo**
**Dirigido por Victoria López López**

**Facultad de Informática**
**Universidad Complutense de Madrid**

**Octubre 2016**

Documento maquetado con TEXIS v.1.0+.

Este documento está preparado para ser imprimido a doble cara.

# Auditoría de Proyectos Big Data, Cloud Computing y Open Data

*Memoria que presenta para optar al título de Graduado en Ingeniería Informática*

**Berta Montes Cobo**
**Dirigido por Victoria López López**

**Facultad de Informática**
**Universidad Complutense de Madrid**

**Octubre 2016**

*A mis padres*

*y*

*a todos aquellos que no creyeron en mí, y me empujaron a luchar más.*

# Agradecimientos

En primer lugar quiero darle las gracias a mi directora, la doctora Victoria López López por brindarme la oportunidad de realizar el trabajo fin de grado con ella sobre una temática que me ha hecho disfrutar, por su guía y su consejo, por la dedicación y la supervisión continua. No puedo olvidar a los profesores que crearon la plantilla Texis, Marco Antonio Gómez Martín y Pedro Pablo Gómez Martín, sin la cual la edición de este documento habría sido más tediosa y mucho más aburrida, y por supuesto a todos aquellos profesores a los que inundé a preguntas en tutorías prácticamente semanales. Gracias por vuestras respuestas teóricas y mil gracias por esas pequeñas frases que te cambian la perspectiva, y te levantan el ánimo cuando la vista se nubla, esas frases que te acompañarán siempre, en concreto recuerdo "En una vida hay muchas vidas, y en la vida se llega a tiempo a todo".

Gracias a los amigos que encontré por el camino, Saúl, Javi, Alberto, Rami, porque siempre han estado para ayudarme a resolver el mundo, por comprender que las navidades y los puentes también son para estudiar y por soportar mis prolongadas ausencias y mi baja disponibilidad.

Gracias a mis padres por ayudarme y animarme en cada tropiezo y en cada victoria de este largo e intenso camino que parecía no tener fin y que sin embargo ahora sé que recordaré con nostalgia, porque sin ellos no sería la mujer en la que me he convertido.

# Resumen

En la actualidad, cada vez son más las empresas y gobiernos que están integrando en sus metodologías de trabajo Cloud Computing y tecnologías Big Data por la versatilidad, flexibilidad, agilidad y seguridad que brindan, con la intención de ser más transparentes, competitivas y eficaces. Además, la transparencia que demanda la sociedad a las instituciones promueve el uso del Open Data, ya que proporciona un acceso libre a gran cantidad de datos con distintos formatos, para su explotación y uso en diversos ámbitos. Como consecuencia, la cantidad de proyectos que usan estas tecnologías es cada vez mayor, manejándose con más frecuencia grandes volúmenes de datos de carácter personal y otros datos generales de distintas fuentes, almacenados en servidores diversos. A pesar de que existen normas para auditar estas tecnologías aún están en fases iniciales y no presentan la visión global que se requiere en proyectos internacionales.

El objetivo principal de este trabajo es analizar el panorama actual, proporcionando así material para aquellos que busquen una referencia a la hora de migrar sus arquitecturas a la nube, introducirse en la filosofía Open Data, como puede ser el Open Science o Open Government entre muchos otros y aumentar sus posibilidades de negocio con tecnologías Big Data, lo que a día de hoy supone una prioridad de negocio.

**Palabras Clave**: Auditoría, Normalización, Estandar, Big Data, Cloud Computing, Datos de Carácter Personal, Open Data, seguridad.

# Abstract

Nowadays, more and more companies and governments are integrating in their work methodologies, consumer models such as Cloud Computing and Big Data technologies because they provide versatility, flexibility, agility and security they provide with the intention of being more transparent, competitive and efficient. In addition, movements such as Open Data are increasingly demanded by society in general, as they promote transparency and provide free access to a large amount of data with different formats, for exploitation and use in various fields. As a result, the number of projects using these technologies is increasing, with large volumes of personal data and other general data from different sources being stored on different servers. Although there are standards for auditing these technologies are still in the initial stages and do not present a global vision as it is required by international projects.

The main objective of this work is to provide an overview of the current scenario, providing material for those who seek a reference when migrating their architectures to the cloud, to be introduced in the Open Data philosophy, such as Open Science or Open Government among many others and increase its business possibilities with Big Data technologies, which today is a business priority.

**Key words**: Audit, Standardization, Standard, Big Data, Cloud Computing, Personal Data, Open Data, Security.

# Índice

# Índice de figuras

# Índice de Tablas

# Capítulo 1

# Introducción

## 1.1. Motivación

La importancia de la realización de este documento radica en dos puntos clave, por un lado se está viendo y cada día más, que este tipo de tecnologías no son una tendencia o una moda pasajera de la industria tecnológica como puede verse en las Figuras 1.1 y 1.2 pues están demostrando ser eficaces y sin duda aún no se está aprovechando al máximo su potencial. Ante un panorama así, es frecuente que el profesional que tenga como objetivo auditar cualquier tipo de proyecto se encuentre cada vez con más asiduidad con una o varias de las tecnologías que aquí se mencionan.

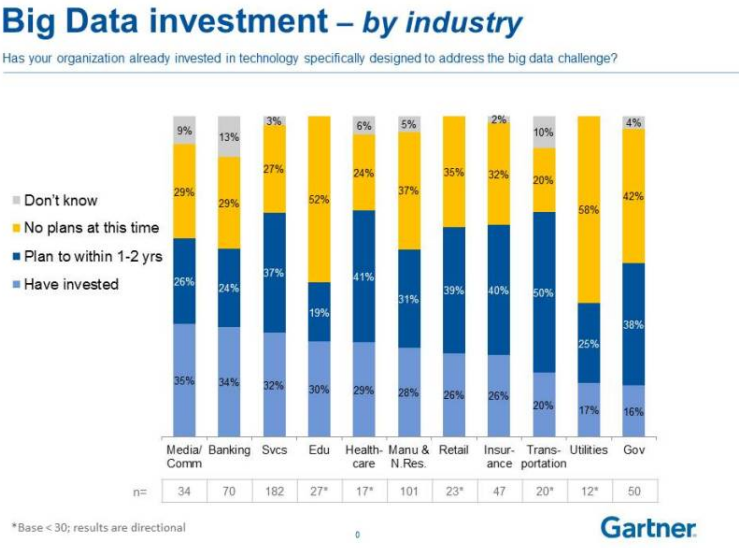


Figura 1.1: Crecimiento del Big Data

Fuente: http://www.datamation.com.ar/cloud-impulsa-el-crecimiento-de-big-data-5343, 2017

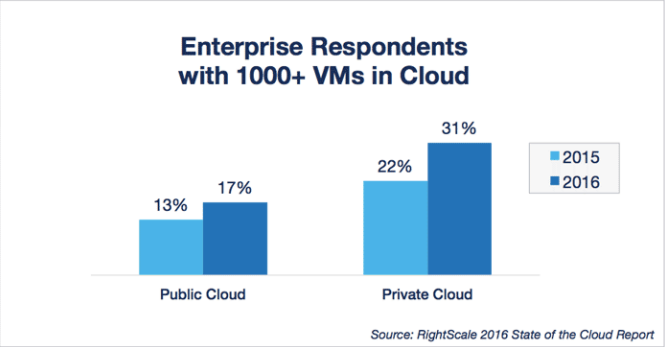


Figura 1.2: Crecimiento del Cloud Computing
Fuente: www.rightscale.com, 2017

Por otro lado, no podemos dejar de mencionar la necesidad que tiene el público en general de comprender estas tecnologías y de saber que aunque el camino está siendo largo, se están haciendo avances para evitar ese "mercadeo de datos" que tanto preocupa a la sociedad en general. Definiendo lo que supone un proyecto Big Data, las dificultades que de manejar un volumen grande de datos, que además crece de manera vertiginosa dado que las fuentes de información son diversas y variadas, (véase las Figuras 1.3, 1.4 y 1.5) o como la computación en la nube se ha convertido en el nuevo modelo de prestación de servicios a través de internet.

## 1.2. Estado del Arte

Todas estas tecnologías aparecieron hace ya unos cuantos años hasta llegar a ser lo que actualmente conocemos. En concreto, en 1944 ya se especulaba con la necesidad futura de trabajar con grandes volúmenes de información no gestionable de forma "tradicional", como se recoge en el artículo “Big Data en la Gestión de Registros de Auditoría” [20]. En dicho artículo se menciona que la necesidad de ampliar las capacidades de procesamiento de datos comienza durante los años ochenta y principios de los noventa del siglo pasado, hasta llegar a nuestros días en los que el almacenamiento en la nube es una opción cada vez más necesaria.

Aún quedan muchos retos que abordar para un mayor aprovechamiento de estas tecnologías, como sistemas obsoletos, incompatibilidad entre estándares y formatos que dificultan la integración de los datos. Sin embargo, se están tomando medidas como puede verse en el Marco Legal que la Unión Europea establece para la protección de datos de carácter personal recogido en la norma “Privacy by Design in Big Data” [21] que analizaremos más

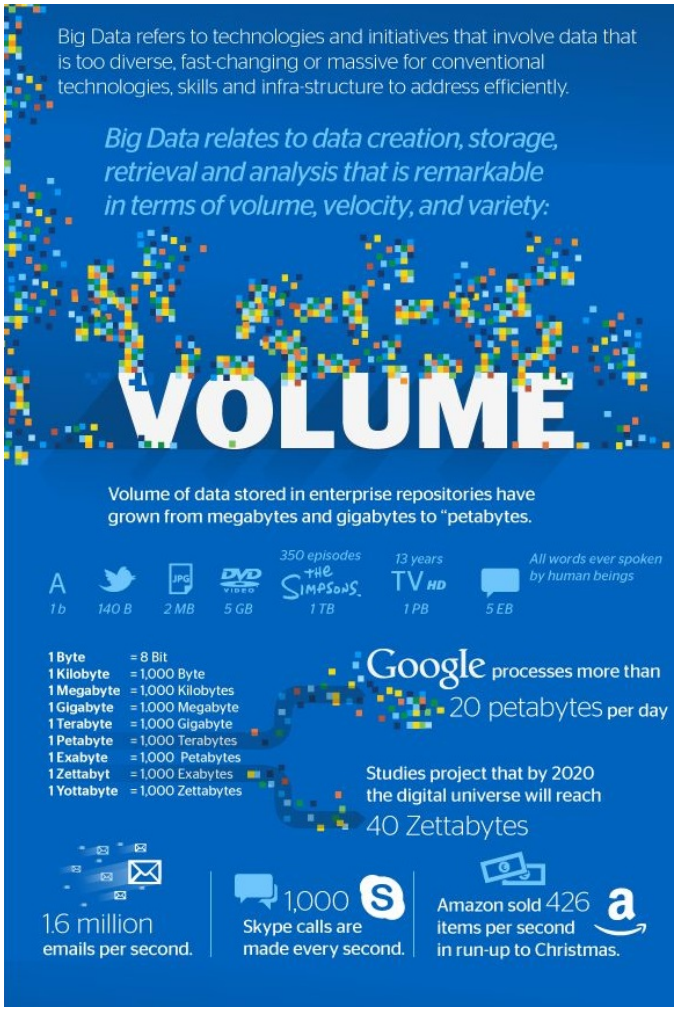


Figura 1.3: Problemas que plantea el Big Data: volumen

Fuente: www.bbvaopen4u.com, 2017

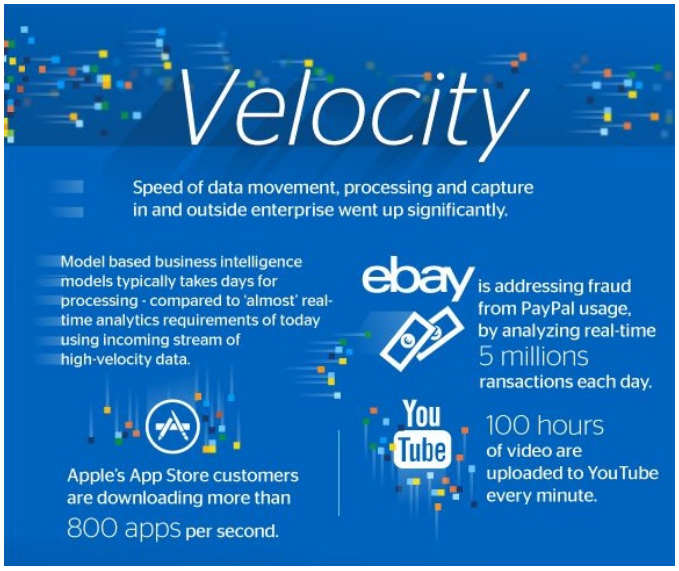


Figura 1.4: Problemas que plantea el Big Data: velocidad

Fuente: www.bbvaopen4u.com, 2017


Figura 1.5: Problemas que plantea el Big Data: variedad

Fuente: www.bbvaopen4u.com, 2017

tarde y otras iniciativas tomadas fuera de España [23].

Por otra parte, la Comisión Europea considera que el Cloud Computing será decisivo en nuestra economía teniendo aún que superar obstáculos como son la protección de datos y la interoperabilidad entre otros aspectos [24].

Por último, pero no por ello menos importante no nos podemos olvidar que tanto el Big Data como el Cloud Computing no representarían lo que actualmente simbolizan sin tener en cuenta movimiento el Open Data, que nació en el 2004 de la mano de la Open Knowledge Foundation [22] está en sus fases iniciales en España, aunque se está extendiendo cada día más, como podemos ver con las iniciativas de algunas comunidades autónomas, o de proyectos llevados a cabo por el gobierno para las administraciones públicas reconocidos por la Unión Europea, como Aporta, nacido para fomentar la cultura Open Data entre la administración pública y la sociedad y promover la reutilización de la información en el sector público. Tampoco nos podemos olvidar de la norma UNE 178301 "Ciudades Inteligentes, Open Data" [15] desarrollada por Aenor para normalizar la publicación de datasets para el sector público como analizaremos en capítulos posteriores.

Uno de los primeros estudios que se llevaron a cabo al realizar esta investigación fue el analizar en detalle la norma ISO 27002 [16] que se comentará con más profundidad en capítulos posteriores, pero que ya nos puede dar cierta información sobre lo que se puede necesitar en proyectos Big Data y/o Cloud Computing. A continuación se muestra un estudio de las cláusulas analizadas. En la Tabla 1.1 se muestra resaltado en color verde las cláusulas específicas para proyectos Big Data, en azul las de Cloud Computing y en rosa aquellas que son aplicables a ambas tecnologías, mientras que en la Figura 1.1 muestra una pequeña estadística de esta misma información.

Tabla 1.1: Índice de la norma ISO/27002

| **Sección** | **Cláusula** |
|---|---|
| 5.1 | 5.1.1 Políticas para la seguridad de la información. |
| | 5.1.2. Revisión de políticas para la seguridad de la información. |
| 6.1 | 6.1.1 Roles y responsabilidades en la seguridad de la información. |

| | |
|---|---|
| | 6.1.2. Segregación de tareas. |
| | 6.1.3. Contacto con las autoridades. |
| | 6.1.4. Contacto con grupos de interés. |
| | 6.1.5. Seguridad de la información en gestión de proyectos. |
| 6.2 | 6.2.1 Política de dispositivos móviles. |
| | 6.2.2. Teletrabajo. |
| 7.1 | 7.1.1 Cribado. |
| | 7.1.2. Términos y condiciones del trabajo. |
| 7.2 | 7.2.1 Gestión de responsabilidades. |
| | 7.2.2. Concienciación, educación y entrenamiento en seguridad de la información. |
| | 7.2.3. procesos disciplinarios. |
| 7.3 | 7.3.1 Terminación o cambio en las responsabilidades del empleado. |
| 8.1 | 8.1.1 Inventario de activos. |
| | 8.1.2. Propietario de los activos. |
| | 8.1.3. Uso aceptable de los activos. |
| | 8.1.4. Devolución de activos. |
| 8.2 | 8.2.1 Clasificación de la información. |
| | 8.2.2. Etiquetado de la información. |
| | 8.2.3. Manejo de activos. |
| 8.3 | 8.3.1 Gestión de medios desmontables. |
| | 8.3.2. Eliminación de medios. |

| | |
|---|---|
| | 8.3.3. Transferencia de medios físicos. |
| 9.1 | 9.1.1 Políticas de controls de acceso.<br>9.1.2. Acceso a redes y servicios. |
| 9.2 | 9.2.1 Registro de usuarios.<br>9.2.2. Provisión de acceso de usuarios.<br>9.2.3. Gestión de los derechos de acceso privilegiado.<br>9.2.4. Gestión de autentificaciones secretas de los usuarios.<br>9.2.5. Revisión de los derechos de acceso de los usuarios.<br>9.2.6. Retirada o ajuste de los derechos de acceso. |
| 9.3 | 9.3.1 Uso de la información de autentificación secreta. |
| 9.4 | 9.4.1 Restricción de acceso a la información.<br>9.4.2. Procedimientos de registro seguro.<br>9.4.3. Sistema de gestión de contraseñas.<br>9.4.4. Uso privilegiado de programas de utilidad.<br>9.4.5. Control de acceso al código fuente del programa. |
| 10.1 | 10.1.1 Políticas de uso en controles criptográficos.<br>10.1.2. Gestión de claves. |
| 11.1 | 11.1.1 Perímetro de seguridad física.<br>11.1.2. Controles de entrada física. |

| | |
|---|---|
| | 11.1.3. Asegurar oficinas, salas e instalaciones. |
| | 11.1.4. Protección contra amenazas externas y medioambientales. |
| | 11.1.5. Trabajar en áreas seguras. |
| | 11.1.6. Zonas de entrega y carga. |
| 11.2 | 11.2.1 Emplazamiento y protección del equipo. |
| | 11.2.2. Utilidades de apoyo. |
| | 11.2.3. Seguridad del cableado. |
| | 11.2.4. Mantenimiento del equipo. |
| | 11.2.5. Eliminación de activos. |
| | 11.2.6. Seguridad de equipos y activos fuera de las instalaciones. |
| | 11.2.7. Borrado seguro o reutilización del equipo. |
| | 11.2.8. Equipo de usuario desatendido. |
| | 11.2.9. Mesa despejada y escritorio limpio. |
| 12.1 | 12.1.1 Procedimientos operacionales y responsabilidades. |
| | 12.1.2. Gestión de cambios. |
| | 12.1.3. Gestión de la capacidad. |
| | 12.1.4. Separación de entornos de desarrollo, pruebas y operacionales. |
| 12.2 | 12.2.1 Controles contra el malware. |
| 12.3 | 12.3.1 Información de respaldo. |
| 12.4 | 12.4.1 Registro de eventos. |
| | 12.4.2. Protección de la información de registro. |
| | 12.4.3. Registro del administrador. |

| | |
|---|---|
| | 12.4.4. Sincronización de relojes. |
| 12.5 | 12.5.1 Instalación de software en sistemas operacionales. |
| 12.6 | 12.6.1 Gestión de vulnerabilidades técnicas.<br>12.6.2. Restricción en la instalación del software. |
| 12.7 | 12.7.1 Controles en auditorías de sistemas de la información. |
| 13.1 | 13.1.1 Controles de red.<br>13.1.2. Seguridad en servicios de red.<br>13.1.3. Segregación en red. |
| 13.2 | 13.2.1 Políticas y procedimientos de transferencia de información.<br>13.2.2. Acuerdos en transferencia de la información.<br>13.2.3. Mensajería electrónica.<br>13.2.4. Acuerdos de confidencialidad o no divulgación. |
| 14.1 | 14.1.1 Análisis de requisitos en sistemas de información.<br>14.1.2. Asegurar servicios de aplicación en la red pública.<br>14.1.3. Protección de transacciones de servicios de aplicaciones. |
| 14.2 | 14.2.1 Políticas de desarrollo seguro.<br>14.2.2. Procedimientos de sistemas de control del cambio. |

| | |
|---|---|
| | 14.2.3. Revisión técnica de aplicaciones después de cambios en la plataforma de operaciones. |
| | 14.2.4. Restricciones en cambios de paquetes software. |
| | 14.2.5. Principios seguros de ingeniería de sistemas. |
| | 14.2.6. Entornos de desarrollo seguro. |
| | 14.2.7. Desarrollo externo. |
| | 14.2.8. Prueba de sistemas de seguridad. |
| | 14.2.9. Prueba de aceptación de sistemas. |
| 14.3 | 14.3.1 Protección de pruebas de datos. |
| 15.1 | 15.1.1 Políticas de seguridad de la información en las relaciones con proveedores. |
| | 15.1.2. Abordar la seguridad dentro de los acuerdos de proveedores. |
| | 15.1.3. Cadena de suministro de tecnología de la información y la comunicación. |
| 15.2 | 15.2.1 Monitorización y revisión de los servicios de proveedores. |
| | 15.2.2. Gestión de cambios en los servicios de proveedores. |
| 16.1 | 16.1.1 Responsabilidades y procedimientos. |
| | 16.1.2. Reportar eventos de seguridad de la información. |
| | 16.1.3. Reportar debilidades en la seguridad de la información. |
| | 16.1.4. Evaluación y decisión sobre los eventos de seguridad de la información. |

| | |
|---|---|
| | 16.1.5. Respuesta a incidentes de seguridad de la información. |
| | 16.1.6. Aprendiendo de los incidentes de seguridad de la información. |
| | 16.1.7. Colección de evidencias. |
| 17.1 | 17.1.1 Planificación de la continuidad de la seguridad de la información. |
| | 17.1.2. Implementación de la continuidad de la seguridad de la información. |
| | 17.1.3. Verificar, revisar y evaluar la continuidad de la seguridad de la información. |
| 17.2 | 17.2.1 Disponibilidad de instalaciones de procesamiento de información. |
| 18.1 | 18.1.1 Identificación de la legislación aplicable y requisitos contractuales. |
| | 18.1.2. Derecho de propiedad intelectual. |
| | 18.1.3. Protección de registros. |
| | 18.1.4. Privacidad y protección de la información de carácter personal. |
| | 18.1.5. Regulación de controles criptográficos. |
| 18.2 | 18.2.1 Revisión independiente de la seguridad de la información. |
| | 18.2.2. Conformidad con las políticas de seguridad y normas. |
| | 18.2.3. Revisión de conformidad técnica. |

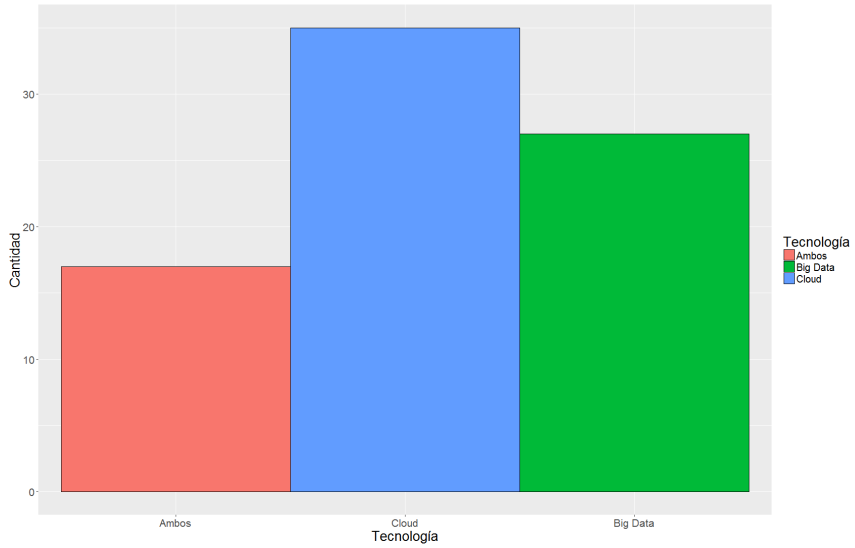


Figura 1.6: Estadística de aplicación de cláusulas

## 1.3. Estructura de este documento

Este documento, consta de cinco capítulos. Además de esta introducción en el segundo capítulo se describe brevemente el concepto de Big Data y se centra en exponer con más detalle las medidas que se están aplicando en referencia a la seguridad de la información en entornos Big Data, así como la normativa europea que existe para la protección de datos de carácter personal recogida en "Privacy by Design in Big Data, an overview of privacy enhancing tecnologies in the era of big data analitycs" [21], también se analiza la norma ISO/27002 [16], además se incluyen propuestas para la auditoría de proyectos con tecnologías Big Data. En el tercer capítulo se expone brevemente los conceptos de Cloud Computing, se hace un breve repaso a las técnicas actuales de anonimización y se analiza la Security Framework for Governmental Clouds [24]. En el cuarto capítulo se describen los principales conceptos de Open Data, y se expone un resumen de la norma española UNE 178301:2015 'Ciudades Inteligentes, Open Data' [15]. Finalmente, en el capítulo quinto se presentan las conclusiones y se mencionan las posibles líneas de trabajo futuro, con el fin de enriquecer y continuar con la investigación realizada. Además, se adjuntan como anexos los formularios para el auditor de los capítulos dos, tres y cuatro a los que también se puede acceder desde internet [3].

# Capítulo 2

# Especificaciones para Auditorías de Proyectos Big Data

*"Without data, you're just another person with an opinion."*

William Edwards Deming.

## 2.1. Introducción

Big Data está cambiando la forma en la que nos relacionamos, abarca diversos ámbitos, desde la forma en la que hacemos negocios, o la manera de afrontar una investigación científica hasta la forma en la que se gobiernan las sociedades. Los datos son recopilados de diversas fuentes, a cualquier hora y en cualquier lugar, siendo el origen más habitual las redes sociales o fichas médicas como muestra la Figura 2.1, con las consecuencias que ello conlleva para la seguridad al tratarse de datos de carácter personal. Por ello las organizaciones están invirtiendo fuertemente en tecnologías Big Data ya que ofrece grandes posibilidades de negocio, y en la ciencia de los datos o data science que se está posicionando como una nueva disciplina científica, proporcionando técnicas, métodos y herramientas para obtener valor de los conjuntos de datos nuevos y de los ya existentes.

La abundancia de datos combinada con la potencia de las técnicas de la ciencia de datos tiene el potencial de mejorar nuestras vidas, haciendo posible la aparición de nuevos productos y servicios. La importancia del "data science" es ampliamente admitida, pero también hay grandes preocupaciones con respecto al uso de los datos. Usuarios y consumidores, están preocupados por un uso irresponsable de los datos, pudiéndose compartir de forma

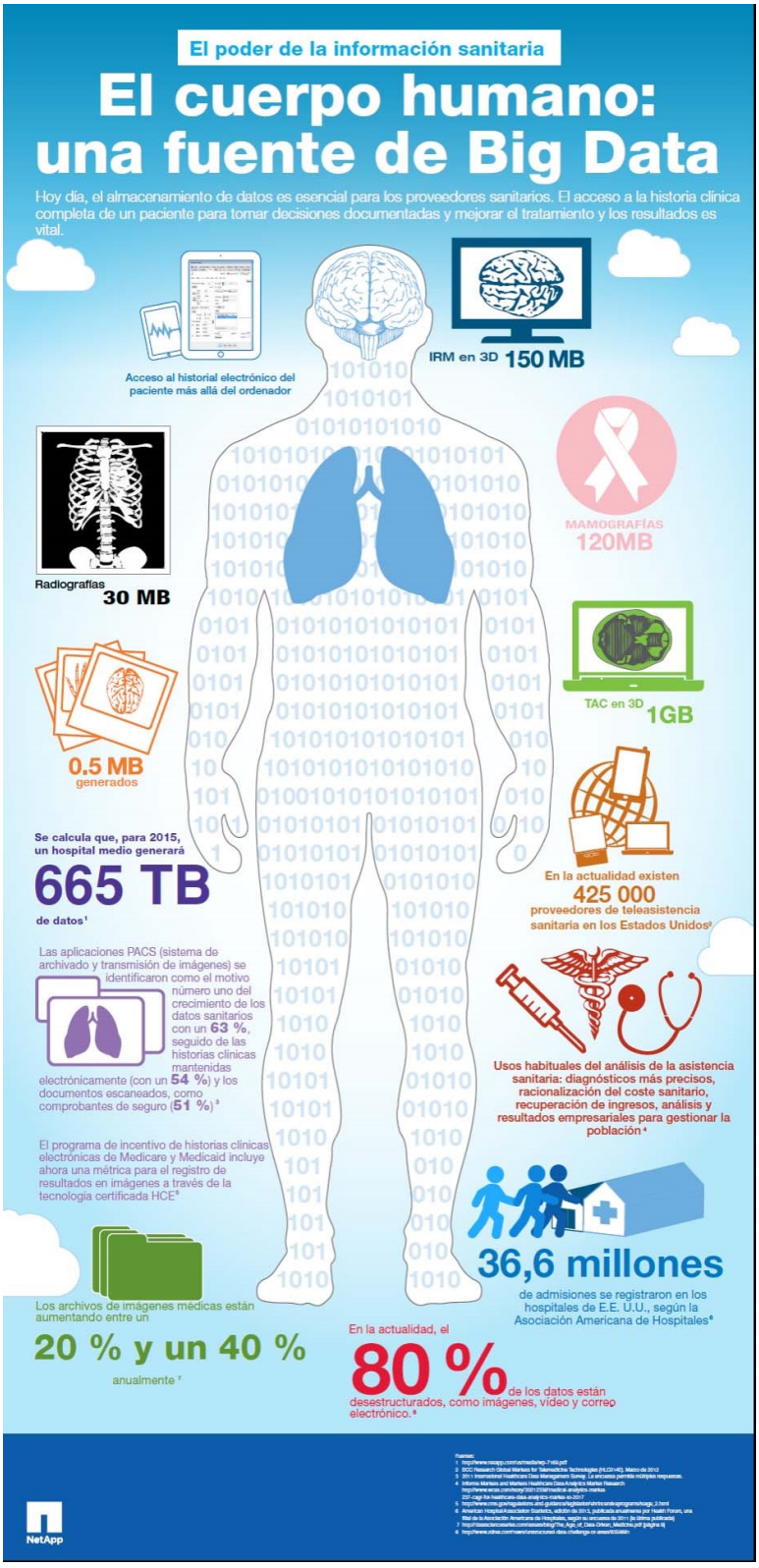


Figura 2.1: Big Data en la medicina

Fuente: www.designinfographics.com, 2017

inintencionada datos confidenciales o que estos sean usados por terceros de manera fraudulenta. Y es que cada etapa en la cadena del data science, (representado en la Figura 2.2) desde los datos en bruto hasta las conclusiones puede crear inexactitudes.

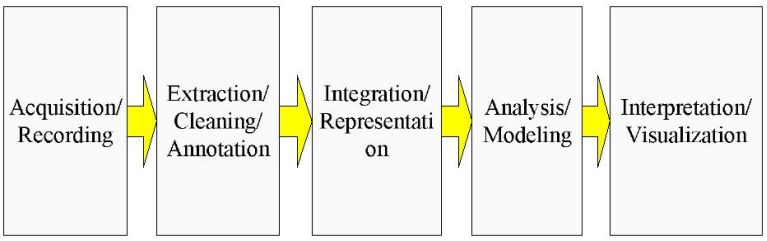


Figura 2.2: Pipeline del data science

Estas preocupaciones pueden conducir a una resistencia del uso a gran escala de los datos y por lo tanto a evitar facilitarlos de manera consciente o a falsearlos en la medida de lo posible, lo que provocaría que fuera imposible obtener beneficio alguno. Es aquí donde radica la importancia de una normalización que garantice la seguridad de los datos en cada una de las etapas del proceso. A este respecto se están tomando medidas como las que podemos ver en el marco legal europeo para la protección de datos de carácter personal recogido en la norma "Privacy by Design in Big Data" [21] que se analiza en el siguiente apartado.

## 2.2. Especificaciones para SGSI

Un Sistema de Gestión de la Seguridad de la Información (SGSI) es un conjunto de políticas que abarcan el diseño, implementación y mantenimiento de un conjunto de procesos y procedimientos que se utilizan para asegurar que una organización pueda obtener el éxito en todas las tareas necesarias, para alcanzar sus objetivos de seguridad de la información respetando la confidencialidad, integridad y disponibilidad de la información.

### 2.2.1. Seguridad Física

Las siguientes cláusulas han sido recogidas de la norma ISO 27002 [16] y analizadas para proyectos Big Data.

*11.1.1 Perímetro de seguridad física*

Se deben definir perímetros de seguridad que sean usados para proteger las áreas que contengan tanto información crítica como sensible, así como las instalaciones donde se procesa. El emplazamiento y la fuerza de cada uno de estos perímetros dependerá de los requisitos de seguridad de los activos dentro del perímetro y del resultado de la evaluación de riesgos. El perímetro de un edificio que contenga instalaciones que procesen información valiosa o sensible, debe ser robusto, entendiendo que todas las puertas exteriores deben estar equipadas con mecanismos de control contra accesos no autorizados y sistemas de detección de intrusos, las puertas y ventanas por tanto deben estar cerradas y se debe tener en cuenta una protección extra para las ventanas que se encuentran al nivel del suelo. Se debería contar con la presencia de una recepción u otra manera de controlar el acceso físico al edificio, además el acceso a las partes del edificio debería estar restringido a personal autorizado. Todas las puertas antiincendios en un perímetro de seguridad deben contar con un sistema de alarma, que debe estar monitorizado y bajo la supervisión de personal autorizado y cualificado. Además, las instalaciones donde se procese la información gestionada por la organización debe estar físicamente separada de aquellas gestionadas por terceros.

*11.1.2 Controles Físicos de entrada*

Se deben proteger las áreas que se consideren necesarias por controles apropiados de entrada que garanticen que sólo el personal autorizado tiene acceso.

*11.1.4 Protección contra amenazas externas y medioambientales*

Se deben aplicar procedimientos contra desastres naturales o ataques intencionados como inundaciones, fuegos o explosiones.

*11.2.1 Protección y emplazamiento físico del equipo*

El equipo debe situarse y protegerse para reducir los riesgos provenientes de amenazas medioambientales y peligros provocados por accesos no autorizados, pudiendo estar situado en un lugar que minimice los accesos innecesarios en áreas de trabajo. Las instalaciones que manejen información sensible deben estar ubicadas con cuidado para reducir el riesgo de que la información pueda ser vista por personal no autorizado durante el uso o de manera accidental.

*11.2.2 Utilidades de apoyo*

El equipamiento debe estar protegido ante fallos eléctricos y otras inte-

rrupciones causadas por fallos en otros sistemas de apoyo. Las instalaciones eléctricas, de gas, agua, ventilación y aire acondicionado deben ser inspeccionadas y probadas regularmente para asegurar su correcto funcionamiento. Así mismo, se debe considerar la posibilidad de tener múltiples suministros con distinto itinerario físico.

*11.2.3 Seguridad en el cableado*

Los cables de alimentación y telecomunicaciones que transporten datos o servicios de información de apoyo deben estar protegidos contra toda interferencia o daño, estas líneas de telecomunicaciones y de alimentación deben ir bajo tierra o sujetas a otras alternativas adecuadas en la medida de lo posible. Los cables de alimentación y los de comunicaciones deben estar separados para evitar interferencias y para los sistemas sensibles o críticos se deben usar blindajes electromagnéticos.

*11.2.4 Mantenimiento del equipo*

El equipo debe ser correctamente mantenido de acuerdo a las recomendaciones y especificaciones del fabricante para asegurar su continuidad, disponibilidad e integridad. Sólo personal autorizado que pueda llevar a cabo un registro total de todos los fallos y de las acciones correctivas que se tomaron para solventarlos debería encargarse del mantenimiento de los equipos. Antes de volver a poner el equipo en funcionamiento una vez realizado el mantenimiento, se debe inspeccionar para asegurarse de que el equipo no está manipulado.

*11.2.5 Eliminación de activos*

El equipo, información, o software no debe ser sacado de su emplazamiento físico o lógico sin una autorización previa, los empleados y usuarios externos que tengan dicha autorización deben ser identificados, además los tiempos para la retirada de activos deben estar establecidos y verificados para poder llevar un registro de cuando se retira un activo y cuando se restaura. La identidad o rol de cualquiera que maneje o use los activos debe ser documentado y esta documentación debe ser devuelta con el equipo, información o software.

### 2.2.2. Seguridad Lógica

*6.1.5 Seguridad de la información en la administración de proyectos*

La seguridad de la información debe abordarse en la gestión del proyecto, independientemente del tipo de proyecto, y debería estar integrada en los métodos de administración de proyectos de la organización, para asegurar que los riesgos de la seguridad de la información son identificados y dirigidos como parte íntegra del mismo. Los métodos de administración de proyectos suelen requerir:

- Los objetivos de seguridad de la información deben estar incluidos entre los objetivos del proyecto.
- La evaluación de un riesgo de seguridad de la información se debe realizar en una etapa temprana del proyecto para poder identificar con tiempo suficiente los controles necesarios para corregir la situación.
- La seguridad de la información debe ser parte de todas las fases de la metodología que se use para desarrollar el proyecto.

Las implicaciones de la seguridad de la información deben ser dirigidas y revisadas regularmente en todos los proyectos. Así, la responsabilidad de la seguridad de la información debe ser definida y asignada a roles específicos, en casos en los que el proyecto incluya tecnología Big Data, esta responsabilidad recaerá sobre el científico de datos.

### *8.2.1 Clasificación de la información*

La información debe ser clasificada en función de sus requisitos legales, valor, criticidad y sensibilidad. Los esquemas de clasificación que se empleen deben incluir criterios que permitan revisar dicha clasificación a lo largo del tiempo así como tener en cuenta las necesidades empresariales de competir o restringir la información, el nivel de protección debe ser evaluado mediante el análisis de la confidencialidad, integridad, disponibilidad, y cualquier otro requisito que se considere necesario. Este esquema debe ser consistente a través de toda la organización, de manera que todo el mundo clasifique la información de igual modo. La creación de grupos de información con el mismo nivel de protección puede ser de utilidad, pero se debe tener en cuenta que la información puede dejar de ser sensible o crítica pasado un tiempo, o por el contrario puede estar clasificada como crítica durante todo su ciclo de vida como ocurre con los datos de carácter personal.

Está cláusula cobra especial relevancia en proyectos Big Data, donde la información proviene de diversas y variadas fuentes, ya que en estos casos se puede dar la situación de manejar datos de carácter personal que precisan de un tratamiento especial.

*8.3.2 Eliminación de medios*

Los dispositivos deben ser borrados de manera segura cuando ya no sean necesario usando procedimientos formales. Estos procedimientos deben estar establecidos para minimizar el riesgo de fuga de información confidencial a personas no autorizadas, en el caso de que se usen para el borrado seguro de los dispositivos que contengan información confidencial deben ser proporcionales al nivel de confidencialidad que tenga la información. Para esto se deberían tener en cuenta los siguientes criterios:

- Los dispositivos que contengan información confidencial deben ser almacenados y borrados de manera segura.
- Deben establecerse procedimientos para identificar los elementos que requieran un borrado seguro.
- Puede resultar más fácil agrupar todos los dispositivos que necesiten ser borrados, en lugar de intentar separar los elementos sensibles.
- Muchas organizaciones ofrecen servicios de recogida y borrado de dispositivos, es recomendable que las empresas externas proporcionen documentación contractual de sus procedimientos.

El borrado de elementos sensibles debe ser registrado con el fin de mantener el trazado de auditoría. Cuando se acumulen dispositivos para el borrado, se debe considerar el efecto de agregación, el cual puede causar que una gran cantidad de datos no sensibles se conviertan en sensibles, este punto tiene una especial relevancia en proyectos Big Data en los que el principio de anonimización no se implemente correctamente, los dispositivos dañados que contengan información sensible pueden requerir una evaluación especial del riesgo para determinar si los elementos deben ser físicamente destruidos en lugar de enviarlos a reparar o descartarlos.

*9.2.5 Repaso de los derechos de acceso de los usuarios*

Se debe revisar de manera frecuente el acceso de los usuarios sobretodo después de ascensos, degradaciones, despidos o cualquier otro cambio significativo.

*10.1.1 Política en el uso de controles criptográficos*

Se debe desarrollar e implementar una política de uso de controles criptográficos para la protección de la información, basándose en la evaluación de riesgos el nivel de protección requerido debe ser identificado teniendo en

cuenta el tipo, fuerza y calidad del algoritmo de encriptación. También es necesario usar técnicas criptográficas para proteger la información transportada en dispositivos móviles o dispositivos desmontables como pueden ser las memorias flash, sin olvidar por otro lado el impacto que puede tener usar información encriptada en controles que se basan en analizar el contenido de estos dispositivos como pueden ser los controles contra la detección de malware. Por último, no podemos olvidar tener en consideración regulaciones y restricciones nacionales que se puedan aplicar al uso de técnicas criptográficas en distintas partes del mundo y a los problemas de flujo transfronterizo de la información encriptada cuando la organización implante sus criptográficos, estos controles pueden ser utilizados para alcanzar diferentes objetivos de seguridad de la información, como puede ser:

- Conformidad: El uso de información encriptada para proteger información crítica o sensible ya sea almacenada o transmitida.
- Integridad/autenticidad: Usar firmas digitales o códigos de autenticación para verificar la autenticidad o integridad de la información sensible o crítica transmitida y/o almacenada.
- No repudio: El uso de técnicas criptográficas para probar la evidencia de la ocurrencia o no de un evento.
- Autentificación: El uso de técnicas criptográficas para autenticar usuarios y otras entidades que soliciten acceso a recursos.

En lo relacionado a proyectos Big Data, no solo es conveniente el uso de controles criptográficos, sino que además deben ser complementados con controles o procedimientos que implementen el principio de anonimización como se verá en la sección 3.3.

#### *12.2.1 Controles contra el malware*

Se deben implementar controles de detección, prevención y recuperación que garanticen la protección contra el malware, así como concienciar a los trabajadores. Estos controles deben detectar el uso de websites maliciosas, y el uso de software no autorizado. Es necesario establecer una política formal para protegerse contra el riesgo asociado con la obtención de archivos y/o software tanto de redes externas como de cualquier otro medio. Se deben realizar revisiones regulares del software y del contenido de los datos de los sistemas que soportan procesos críticos, prestando atención a la instalación y frecuencia de actualización del software que detecta y repara las acciones del malware. Si durante estas revisiones se detecta la presencia de cualquier fichero no autorizado, éste debe ser formalmente investigado. Una rutina básica podría incluir:

- Escaneo de archivos recibidos por red y/o almacenados en disco en busca de malware antes de cualquier uso.
- Escaneo de archivos adjuntos en correo electrónico. Este escaneo debería hacerse en distintas partes (servidores mail, sobremesa, y cuando se conecte a la red de la empresa).

Se aconseja contar con un experto en seguridad informática entrenado en la recuperación de datos tras un ataque malware. Además es necesario preparar un plan de continuidad de negocio para la recuperación de ataques por malware incluido todos los datos y software para el backup. Se recomienda aislar los entornos donde el daño por un ataque pueda resultar catastrófico.

*12.3.1 Información de respaldo*

Se debe establecer una política de respaldo que defina y establezca los requisitos de la empresa en cuanto a respaldo de la información, del software y de los sistemas, para ello se debe contar con instalaciones de respaldo adecuadas que garanticen que toda la información esencial así como el software se puede recuperar después de un ataque o un fallo en los dispositivos, además las copias de seguridad han de ser probadas regularmente para asegurar su funcionamiento en caso de incidente. Cuando se diseñe el plan de respaldo se debería tener en cuenta que:

- Deben producirse registros exactos y completos de las copias y de los procedimientos documentados de restauración.
- La extensión y frecuencia de los respaldos debe reflejar los requisitos de la organización, los requisitos de seguridad de la información involucrada y la criticidad de la información.
- Las copias de respaldo deben ser almacenadas en un emplazamiento remoto, a una distancia suficiente como para que no sufra los daños de un posible desastre en el edificio principal.

A la información de respaldo se le debe dar el apropiado nivel de seguridad física, y cuando la confidencialidad sea necesaria, se debe proteger la información de respaldo con las medidas oportunas de encriptación y anonimización.

*13.1.1 Controles de Red*

Las redes deben ser controladas y gestionadas para proteger la información de los sistemas y aplicaciones. Se deben implementar controles que

garanticen la seguridad de la información en red y la protección de los servicios conectados desde accesos no autorizados. En particular:

- Se deben establecer y determinar procedimientos y responsabilidades para la administración de los equipos en red.
- La responsabilidad operacional de las redes debe separarse de las operaciones informáticas cuando proceda.
- Se deben tomar controles especiales para salvaguardar la confidencialidad e integridad de los datos al pasar por redes públicas o inalámbricas y para proteger los sistemas y aplicaciones conectados. También se pueden requerir controles especiales para mantener la disponibilidad de los servicios en red y ordenadores conectados.

Las actividades de gestión deben ser estrictamente coordinadas tanto para optimizar el servicio o la organización como para asegurar que los controles se apliquen consistentemente a través de la infraestructura de procesamiento de la información. Los sistemas de la red deben ser autentificados y la conexión de los sistemas a la red debe restringirse [17].

*14.2.8 Pruebas de Seguridad del Sistema*

Se deben realizar pruebas de seguridad, tanto los sistemas nuevos como los actualizados requieren pruebas de verificación durante el proceso de desarrollo, incluida la preparación de un plan de actividades y pruebas bajo un rango específico de condiciones. Para desarrollos in-house, dichas pruebas deben llevarse a cabo por el equipo de desarrollo, además deberían realizarse pruebas de aceptación independientes para asegurarse de que el sistema funciona como se espera tanto en arquitecturas in-house como en arquitecturas cloud.

*16.1.2 Reportar eventos de seguridad de la información*

Los eventos de seguridad de la información deben ser comunicados por los canales administrativos apropiados tan rápido como sea posible. Todos los empleados deben ser conscientes de su responsabilidad para reportar eventos de seguridad tan rápido como puedan. Además deben conocer los procedimientos disponibles para comunicar los eventos de seguridad de la información y los puntos de contacto a los que cada tipo de evento debe ser reportado. Algunas situaciones a tener en cuenta para reportar eventos de seguridad de la información son:

- Controles de seguridad ineficientes.
- Brecha en la integridad y/o confidencialidad de la información.
- Errores humanos.
- No conformidad con las políticas o guías.
- Brecha física de seguridad.
- Cambios no controlados en el sistema.
- Mal funcionamiento del software y/o hardware.
- Violación de accesos.

*16.1.3 Informar sobre las debilidades en la seguridad de la información*

Los empleados y contratistas que usen los sistemas de información de la organización están obligados a reportar cualquier debilidad de seguridad de la información observada o sospechada en el sistema y/o en los servicios. Todos los empleados y contratistas deben informar de estos asuntos, tan rápido como sea posible con el fin de prevenir incidentes graves de seguridad. El mecanismo para informar de dichos incidentes debe ser fácil y accesible. Además debe poder ser comprobado por el auditor.

*16.1.6 Aprendiendo de los incidentes de seguridad de la información*

Debe haber mecanismos que permitan cuantificar y monitorizar los tipos, volumen y costes de los incidentes de seguridad, la información obtenida de los análisis debe usarse para reducir el impacto de futuros sucesos, identificar incidentes recurrentes o de alto impacto. Además toda esta información debe estar convenientemente documentada para proporcionar evidencias de auditoría.

*18.1.3 Protección de archivos*

Se debe tener en cuenta la forma en la que la organización clasifica sus archivos y el período de tiempo durante el que los almacena a la hora de seleccionar el medio en el que va a guardar la información para protegerla de accesos no autorizados o descargas ilegales, si los archivos son digitales cualquier clave criptográfica asociada debe almacenarse de manera segura para poder habilitar el descifrado si fuera necesario. Por lo que se considera recomendable establecer directrices sobre la conservación, manipulación y eliminación de los registros, así como un calendario que determine el período de tiempo durante el que pueden ser retenidos los archivos.

### 2.2.3. Datos de carácter personal

*18.1.4 Privacidad y protección de la información personal*

Se debe desarrollar e implementar una política para la protección y privacidad de la información de carácter personal y todas las personas involucradas en el procesamiento de información personal deben conocerla.

*18.2.1 Revisión independiente de la seguridad de la información*

El enfoque de la organización para gestionar la seguridad de la información y su implementación debe revisarse a intervalos planificados o cuando se produzcan cambios significativos. Debe ser la dirección de la organización quien tome la iniciativa de solicitar la revisión, dicha revisión debe ser llevada a cabo por individuos ajenos a la organización y/o independientes del área a revisar. Los resultados de está revisión deben ser almacenados y notificados a la parte directiva que los solicitó. Si la revisión identifica que la política implementada por la organización para la seguridad de la información es inadecuada, la dirección debe considerar las acciones correctivas que le sean recomendadas.

#### 2.2.3.1. Marco Legal europeo para la protección de datos de carácter personal

El actual marco jurídico para evaluar las cuestiones de privacidad planteadas por la analítica Big Data en la Unión Europea está compuesto por la Directiva 95/46/EC y la Directiva 2002/58/EC (modificada por la Directiva 2009/136/EC). Cuando los datos sean correctamente anonimizados es decir, hayan sido procesados de forma que no puedan ser usados para identificar a una persona, es posible que que el procesamiento de los mismos pueda salirse del marco legal de la Unión Europea, a tales escenarios hace referencia la Directiva 95/46/EC. En el contexto del Big Data el responsable de datos o 'data controllerés crucial para la protección de los mismos ya que determina el propósito del procesamiento, por ello tiene las siguientes obligaciones:

En primer lugar, el responsable de datos debe asegurarse de que los datos son recogidos de manera justa, legal (recogidos para fines específicos y legítimos) y se cuenta con el consentimiento previo del individuo que los ha entregado de manera voluntaria para procesarlos, habiendo sido previamente informado de manera específica y no ambigua. En segundo lugar, los datos procesados serán los estrictamente necesarios para el propósito por el cual se recolectaron, a esto se le conoce como "Principio de Minimización de Datos". También están obligados a almacenar los datos personales pro-

cesados exclusivamente el tiempo mínimo necesario para el propósito por el cual fueron recogidos, pasado este período deberán ser borrados. El artículo 5 apartado 3 de la Directiva 2002/58/EC se aplica a las situaciones en las que los interesados en Big Data almacenan o tienen acceso a la información contenida en los dispositivos de los usuarios. Este abastecimiento exige que el usuario consienta que se almacene o acceda a su dispositivo para obtener la información. Por otro lado, hay ciertos requisitos de transparencia que también recaen sobre el responsable de datos, éste debe proporcionar su identidad, los propósitos del procesamiento, los contenedores de los datos. La disponibilidad y claridad de esta información es imprescindible para validar el consentimiento de los usuarios, y es una herramienta fundamental para lidiar con cualquier posible discrepancia que pudiera ocurrir entre los usuarios y los proveedores de servicios. El articulo 17 de la Directiva de protección de datos califica al responsable de datos como total responsable de la seguridad de los mismos, dotándole de la obligación de implementar las medidas apropiadas para proteger los datos personales, en el contexto Big Data, esto significa principalmente la implementación de controles para limitar el acceso a personas no autorizadas.

Los propietarios de los datos también tienen derechos en el entorno Big Data, en particular cualquier interesado tiene derecho a solicitar al responsable de datos y obtener de él sus datos, conocer la lógica utilizada en el tratamiento de los mismos o revocar cualquier consentimiento previo. El ejercicio de estos derechos no debería suponer ninguna carga para el titular.

#### 2.2.3.2. Propuesta para el reglamento general de protección de datos

La Directiva 95/46/EC se refina con el fin de hacer frente al nuevo contexto tecnológico, introduciendo también garantías para los titulares de los datos y nuevos derechos de usuario, a este respecto hay muchos elementos interesantes para señalar en relación al Big Data. La transparencia toma protagonismo, los datos de contacto del responsable, las bases legales del procesamiento (incluidos los intereses legítimos específicos perseguidos por el responsable del tratamiento o por un tercero), y la existencia del derecho a la portabilidad de datos, deben estar recogidos en un documento que le sea facilitado al usuario de una manera estructurada y en un formato legible para una máquina, y tiene derecho a dársela a otro responsable de datos. El derecho de borrar se verá reforzado con la introducción de lo que se llama "Derecho al Olvido" con el objetivo de propiciar una actitud más responsable y cautelosa hacia la divulgación pública de datos. En particular, este nuevo derecho prevé la obligación para un responsable de datos, de tomar medidas razonables para informar a otros responsables de procesamiento, de que el

interesado ha solicitado el borrado de cualquier enlace, copia o reproducción de los datos de carácter personal.

Se introducen nuevos mecanismos de control para hacer frente a la complejidad de la cadena de valor, así dos o más controladores de datos determinan conjuntamente los fines y medios del tratamiento de los datos de carácter personal. Se deberá entonces determinar sus respectivas responsabilidades y decidir cuál de los dos actuará como enlace con los usuarios para que estos puedan ejercer libremente sus derechos.

Con el fin de que el usuario se encuentre seguro respecto al tratamiento de sus datos y evitar así la sensación de "mercadeo de datos" de la que hablábamos antes, el responsable de datos debe asegurarse de que la reutilización de éstos no es incompatible con el propósito original que provocó la colecta. Por lo tanto, la transparencia con los individuos y la utilización de herramientas válidas para expresar sus elecciones son las únicas maneras de llevar a cabo una reutilización de datos eficiente y responsable.

## 2.3. Recomendaciones Generales

Para definir como la privacidad desde el diseño se integra completamente en el análisis Big Data mediante el uso de herramientas y procesos de apoyo, los responsables de datos y analistas deben trabajar conjuntamente con las autoridades de protección de datos. Para poder sacar el máximo partido al análisis Big Data es necesario acceder con seguridad a la información, la analítica descentralizada puede ser beneficioso en muchos casos, sobretodo cuando el tipo de análisis es conocido y los datos vienen de muchas fuentes diferentes.

Desarrollar una política de privacidad que respete las obligaciones legales es clave en cualquier escenario donde se procesen datos, Big Data introduce un retos adicionales, la cadena de responsables de datos e intercambio de información, y el hecho de que ciertos requisitos de privacidad de un responsable de datos pueden no ser respetados por oto. Involucrar activamente a los usuarios finales para que demanden transparencia y control es esencial, de esta forma toman conciencia de su privacidad. En cuanto a Big Data se refiere, los mecanismos tradicionales de consentimiento fallan desde el punto de vista de la transparencia, se habla de privacidad pero la propia idea de "consentimiento" necesita ser reforzada, para que deje de ser interpretada como una carga o un "riesgo" que deben correr los usuarios. Ya hay muchas herramientas para mejorar la privacidad online y en los dispositivos móviles como tecnologías antitracking, de encriptación y herramientas de comparti-

ción de archivos de manera segura, que pueden ofrecer un apoyo valioso.

## 2.4. Recomendaciones Tecnológicas para la Privacidad

En Europa se ha reformado el marco legal para incluir nuevos derechos y obligaciones asociados con el Big Data, especialmente en relación con el establecimiento del responsable de tratamiento, eliminación de datos en la red y la transferencia de datos fuera de la Unión Europea. Dado que la implementación de obligaciones legales no es sencilla y la regulación es de suma importancia, la protección de datos de carácter personal debe apoyarse también en el crecimiento tecnológico, por ello la tecnología para el Big Data debe apoyarse en la tecnología para la privacidad.

Con el fin de obtener resultados provechosos los datasets disponibles en Big Data deberían ser tan amplios como fuera posible, sin embargo la minimización de datos y las restricciones que imponen los datos de carácter personal son una parte fundamental de la privacidad desde el diseño. Encontrar un equilibrio saludable entre estos dos enfoques es uno de los mayores retos que tiene el Big Data, minimizar la recogida de datos permitiendo al mismo tiempo un contenido rico y útil para el análisis. A menudo la información más relevante para el Big Data es la que deja el usuario al usar la tecnología día a día sin ser apenas consciente de ello. La unión de información de diferentes fuentes es una parte esencial del análisis Big Data, así el principio de "Privacidad Desde el Diseño" se compone de varias etapas que deben respetarse para asegurar la integridad de los datos durante todo el proceso. La Figura 2.3, muestra una visión general de las estrategias de privacidad desde el diseño que se analizan a continuación.

Estrategias de privacidad en la cadena de valor del Big Data.

- Adquisición de datos
  - Minimizar: Cada responsable que esté recolectando datos tiene que definir qué datos personales se necesitan y cuáles no para el propósito del proceso, incluyendo también los periodos de retención de los mismos. El estudio del impacto de privacidad es una herramienta valiosa para determinar las necesidades de procesamiento exactos, y limitar los datos a los estrictamente necesarios para el fin determinado.
  - Agregar: Cuando el análisis estadístico de fuentes distribuidas de

| | PRIVACY BY DESIGN STRATEGY | DESCRIPTION |
|---|---|---|
| 1 | Minimize | The amount of personal data should be restricted to the minimal amount possible (data minimization). |
| 2 | Hide | Personal data and their interrelations should be hidden from plain view. |
| 3 | Separate | Personal data should be processed in a distributed fashion, in separate compartments whenever possible. |
| 4 | Aggregate | Personal data should be processed at the highest level of aggregation and with the least possible detail in which it is (still) useful. |
| 5 | Inform | Data subjects should be adequately informed whenever processed (transparency). |
| 6 | Control | Data subjects should be provided agency over the processing of their personal data. |
| 7 | Enforce | A privacy policy compatible with legal requirements should be in place and should be enforced. |
| 8 | Demonstrate | Data controllers must be able to demonstrate compliance with privacy policy into force and any applicable legal requirements. |

Figura 2.3: Estrategias de privacidad desde el diseño
Fuente: Privacy by Design in Big Data [21]

datos de carácter personal no sea necesario, y la colección de información anonimizada sea suficiente. La anonimización local es la solución prominente que permite al responsable eliminar toda la información personal antes de liberarla para el análisis.

- Esconder: En muchos casos la información acerca del individuo puede ser recogida sin que éste se de cuenta (busquedas en internet/ comportamiento online ...) Hay tecnologías como el cifrado, el enmascaramiento de la identidad y herramientas para compartir archivos de manera segura que potencian la intimidad y apoyan la privacidad en internet y en dispositivos móviles.
- Avisar: Los individuos deben ser informados apropiadamente acerca de la recopilación de sus datos de carácter personal para el análisis del Big Data. Aunque la etapa de recolección de datos es la más importante para el individuo en cuanto a tomar una decisión acerca del uso de sus datos, estas notificaciones deberían estar disponibles para los individuos en todo el proceso del Big Data y no solo durante la recoleta de datos.
- Control: Herramientas en las que se puede elegir *No* deberían ser ofrecidas al individuo durante todo el proceso.

- Análisis de datos
  - Agregar: Una de las técnicas prominentes en el análisis de datos

es la anonimización.

- Esconder: Otra técnica muy importante para preservar la privacidad es la encriptación, la encriptación en búsquedas, la encriptación homomórfica y la computación multiparte segura son tecnologías prometedoras en este campo con un gran interés para la comunidad investigadora.

- Almacenamiento
  - Esconder: medidas de seguridad como el control de acceso granulado y la autentificación son esenciales en la protección de datos de carácter personal en las bases de datos. Tecnologías como el Control de Acceso basado en atributos pueden ser mucho más escalable en Big Data, ofreciendo políticas de control de acceso de grano fino.
  - Separar: control de acceso y técnicas de encriptación.

- Uso de datos
  - Agregar: preservar la privacidad de publicación y recuperación de datos se basa normalmente en la anonimización con el fin de prevenir la inferencia de datos personales. Las cuestiones relacionadas con la procedencia de los datos en el curso de la toma de decisiones basadas en el Big Data es otro tema de interés sobre todo en cuanto a la credibilidad y el nivel de agregación de los metadatos a fin de evitar la identificación de los individuos.

Por encima de las estrategias de diseño mencionados, los responsables de los datos tienen que tener en cuenta las obligaciones legales subyacentes, en particular en relación con los principios de privacidad y la base jurídica. La Figura 2.3 proporciona una visión general de las estrategias de privacidad desde el diseño aplicadas a cada una de las fases de la cadena de valor del Big Data.

Por otro lado, otras instituciones también se están planteando nuevos retos con respecto al Big Data como el RDS (Responsable Data Science), integrado por grupos de investigadores de múltiples disciplinas, este grupo situado en Holanda, se plantea los retos del Big Data desde otra perspectiva haciendo especial hincapié en la responsabilidad y centrándose en la justicia, la exactitud, la confidencialidad y la transparencia para alcanzarla.

El programa que plantean sigue estos cuatro pasos:

- Permitir y garantizar el uso responsable de los datos sin inhibir el poder de la ciencia de datos.

| | BIG DATA VALUE CHAIN | KEY PRIVACY BY DESIGN STRATEGY | IMPLEMENTATION |
|---|---|---|---|
| 1 | Data acquisition/collection | MINIMIZE | Define what data are needed before collection, select before collect (reduce data fields, define relevant controls, delete unwanted information, etc), Privacy Impact Assessments. |
| | | AGGREGATE | Local anonymization (at source). |
| | | HIDE | Privacy enhancing end-user tools, e.g. anti-tracking tools, encryption tools, identity masking tools, secure file sharing, etc. |
| | | INFORM | Provide appropriate notice to individuals – Transparency mechanisms. |
| | | CONTROL | Appropriate mechanisms for expressing consent. Opt-out mechanisms. Mechanisms for expressing privacy preferences, sticky policies, personal data stores. |
| 2 | Data analysis & data curation | AGGREGATE | Anonymization techniques (k-anonymity family, differential privacy). |
| | | HIDE | Searchable encryption, privacy preserving computations. |
| 3 | Data storage | HIDE | Encryption of data at rest.<br>Authentication and access control mechanisms.<br>Other measures for secure data storage. |
| | | SEPARATE | Distributed/ de-centralised storage and analytics facilities. |
| 4 | Data use | AGGREGATE | Anoymisation techniques.<br>Data quality, data provenance. |
| 5 | All phases | ENFORCE/ DEMONSTRATE | Automated policy definition, enforcement, accountability and compliance tools. |

Figura 2.4: Estrategias de privacidad desde el diseño en la cadena de valor del Big Data

Fuente:Privacy by Design in Big Data [21]

- Abstraer la tecnología para garantizar la equidad, la exactitud, la confidencialidad y la transparencia desde el diseño.

- Reunir a los mejores investigadores de disciplinas clave como minería de datos, recuperación de la información, derecho, machine learning, procesamiento del lenguaje natural, seguridad, estadística

- Crear una plataforma multidisciplinaria centrada en los desafíos relacionados con el Big Data y la ciencia de datos.

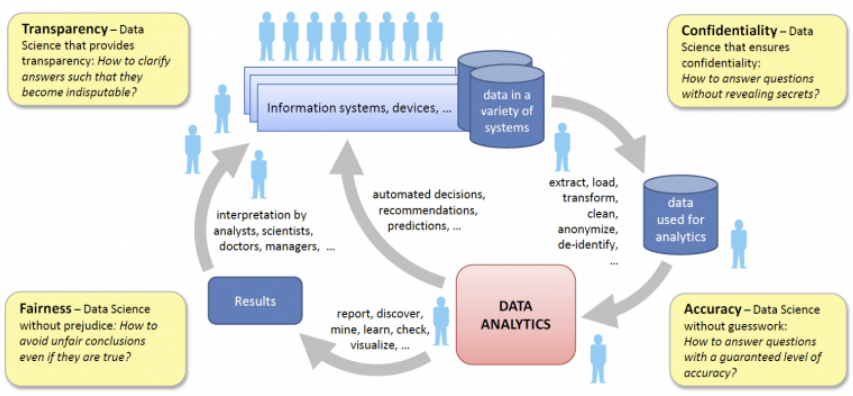


Figura 2.5: Pipeline

Fuente: www.responsibledatascience.org, 2017

La Figura 2.5 proporciona una visión de conjunto de un pipeline típico en data science. La creación de datos desde diversos ámbitos de la sociedad para que posteriormente sean recolectados y a menudo 'cocinados', es decir, extraer, transformar y/o de identificar la información antes de que sea usada en el análisis. Los resultados del análisis incluyen modelos (arboles de decisiones), decisiones automatizadas, predicciones y recomendaciones y los resultados deben ser interpretados por las partes interesadas.

El RDS centra sus esfuerzos en FACT (Fairness, Accuracy, Confidentiality and Transparency) es decir, Justicia, exactitud, confidencialidad y transparencia reconociendo así las preocupaciones de la sociedad. La ciencia responsable de datos, está impulsada por el ideal de incorporar valores y aspectos sociales y éticos al convertir los datos en un valor cuantificable.

- Fairness : Data science sin prejuicios o como evitar conclusiones injustas aunque sean verdad.
- Accuracy: Data science sin conjeturas o como responder con cierto nivel de exactitud.
- Confidentiality: Data science que asegure la confidencialidad o cómo contestar preguntas sin revelar secretos.
- Transparency: Data science que proporcione transparencia o como clarificar respuestas de tal manera que se conviertan en indiscutibles.

Estas cuatro trayectorias están claramente relacionadas y todas contribuyen a una visión global para hacer posible una data science responsable.

# Capítulo 3

# Especificaciones para Auditorías de Proyectos Cloud Computing

*"Lo interesante del Cloud Computing es que estamos redefiniendo el término para incluir en él todo lo que ya hacemos"*

Larry Ellison.

## 3.1. Introducción

Las tecnologías cloud tiene el potencial de mejorar la colaboración entre personas y/o empresas y reducir costes a través de una computación más eficiente que permite la escalabilidad. Pero además Cloud Computing acelera la necesidad de redefinir el concepto de "perímetro de seguridad", las distintas modalidades de despliegue que puede ofrecer Cloud (IaaS, PaaS, SaaS) deben ser pensadas tanto en el contexto de la seguridad física como en el de la seguridad lógica, teniendo en cuenta la diversidad de escenarios y usuarios finales a los que se puede aplicar una solución Cloud. En su mayor parte los controles de seguridad para Cloud Computing no son diferentes de los controles de seguridad de cualquier otro entorno TI como veremos a continuación, muchas de las cláusulas que se aplican al Big Data, también son válidas para Cloud Computing, sin embargo la mayor dificultad que presenta Cloud Computing en cuanto a seguridad se refiere, está en abstraerse de la infraestructura y en la falta de visibilidad de muchos de los controles habituales.

## 3.2. Especificaciones para SGSI

Las siguientes cláusulas han sido recogidas de la norma ISO 27002 [16] y analizadas para proyectos Cloud Computing.

### 3.2.1. Seguridad Física

Las consideraciones a tener en cuenta respecto de la seguridad física son las mismas que las aplicadas en la misma sección del capítulo 2. En la Tabla 1.1 puede verse más claramente que cláusulas se comparten por ambas tecnologías, tanto para seguridad física como para seguridad lógica.

### 3.2.2. Seguridad Lógica

*6.2.1 Políticas de los dispositivos móviles*

Se debe adoptar políticas de seguridad que controlen los riesgos inherentes a los dispositivos móviles, estas políticas son de especial interés en relación a proyectos con tecnologías Cloud Computing ya que uno de los puntos fuertes de esta tecnología es la posibilidad de acceder a la información desde cualquier parte. Para ello se debe tener en cuenta:

- Registro de dispositivos móviles.
- Requisitos de protección física.
- Restricciones en la instalación del software.
- Requisitos para versiones software y para la aplicación de parches.
- Restricciones de conexión a los servicios de información.
- Control de acceso.
- Técnicas criptográficas.
- Técnicas de anonimización.
- Protección contra el malware.
- Desconexión, borrado o bloqueo remoto.
- Copias de respaldo de la información.
- Uso de servicios web y aplicaciones.

Los procedimientos deben considerar la posibilidad de robo o pérdida de los dispositivos móviles y el uso de los mismos en el ámbito privado de los empleados si fuera necesario.

*9.2.4 Gestión de la autentificación secreta de los usuarios*

La asignación de autentificaciones secretas se debe controlar teniendo en cuenta los siguiente criterios:

- Los usuarios deben firmar una cláusula que formará parte de las obligaciones del empleado para mantener en secreto su autentificación personal.
- Los empleados deben disponer de una autentificación provisional que les fuerce a cambiarla si son ellos los responsables de mantener su propia autentificación.
- Las autentificaciones temporales se deben entregar de una manera segura, evitando terceras partes o correos electrónicos desprotegidos, además deben ser únicas para el individuo.

*9.3.1 Uso de información secreta de autentificación*

Los usuarios deben seguir las prácticas de la organización y estar informados de que :

- La información de autentificación tiene que ser confidencial y no debe ser divulgada a terceras personal incluidas aquellas que tengan mayor autoridad o rango dentro de la empresa.
- Se debe evitar el uso de papeles o ficheros electrónicos para anotar la autentificación, a menos que éstos puedan ser almacenados de forma segura y la manera en que se almacenan haya sido aprobada.
- Se debe cambiar la autentificación siempre que se sospeche de que ésta haya podido ser comprometida.
- Se deben seleccionar contraseñas que:
  - Sean fáciles de recordar.
  - No estén basadas en nada que alguien cercano pudiera adivinar, como el numero de teléfono, el nombre o fechas de cumpleaños.
  - No sean vulnerables a ataques de fuerza bruta.
  - Incluyan números y símbolos.

Además si la contraseña es temporal, se debe cambiar la primera vez que se inicie sesión y no se debe usar la misma contraseña en ámbitos laborales y personales.

*9.4.3 Sistema de gestión de contraseñas*

El sistema de gestión de contraseñas tiene que asegurar la calidad de las mismas, para ello debe:

- Permitir que los usuarios cambien sus contraseñas y hagan modificaciones si se comete un error en el proceso.
- Forzar a los usuarios a seleccionar contraseñas de calidad.
- No debe mostrar la contraseña en la pantalla cuando se introduzca.
- Llevar un registro de contraseñas ya usadas para evitar repeticiones.
- Separar los ficheros de información de la aplicación de los ficheros de contraseñas.
- Almacenar y transmitir contraseñas de manera segura.

*11.2.7 Eliminación segura o reutilización del equipo*

Se deben verificar todos los elementos del equipo para asegurarse de que los datos confidenciales y el software se han eliminado de forma segura antes de su eliminación o reutilización, con los dispositivos que contengan información sensible se deben emplear técnicas que garanticen que dicha información no es accesible o destruirlos físicamente si no fuera posible obtener esta garantía.

*12.1.3 Gestión de la capacidad*

Se debe monitorizar el uso de los recursos con la finalidad de realizar predicciones de los requisitos de capacidad necesaria para asegurar el funcionamiento del sistema en situaciones de estrés y evitar cuellos de botella. Para ello se puede ajustar la capacidad controlando la escalabilidad, borrando información obsoleta, optimizando horarios y procesos. Esta cláusula es importante para proyectos Cloud Computing, ya que precisamente con esta tecnología se paga por el servicio prestado/usado.

*13.1.2 Seguridad de los servicios de red*

Se deben identificar todos los servicios de red, ya sean servicios internos o externos para añadirse a los acuerdos, que deben incluir el derecho a la

auditoria y la capacidad del proveedor de servicios de red para gestionar los mismos de manera segura. Se recomienda una supervisión periódica para identificar las medidas de seguridad necesarias de determinados servicios, así como garantizar que los proveedores de servicio de red apliquen estas medidas.

*13.2.1 Políticas y procedimientos de transferencia de la información*

Deben establecerse políticas, procedimientos y controles para proteger la transferencia de información, para ello se deben considerar:

- Procedimientos para proteger la información transmitida de la intercepción, destrucción, copia, modificación y/o enrrutamiento incorrecto.
- Procedimientos para la detección y protección contra el malware que pueda ser transmitido a través de las infraestructuras de comunicación.
- Procedimientos para la protección de información sensible que está en forma de archivo adjunto.
- Políticas que describan el uso aceptable de las instalaciones de comunicación.
- Uso de técnicas criptográficas y de anonimización que protejan la confidencialidad, integridad y autenticidad de la información.
- Controles y restricciones asociados con el uso de instalaciones de comunicación.

*15.1.3 Cadena de suministro de tecnología de la información y las comunicaciones*

Se deben incluir requisitos asociados con la tecnología de comunicación y la cadena de suministro de productos en los acuerdos con proveedores, para abordar los riesgos de seguridad de la información. Las prácticas de seguridad, deben propagarse a través de toda la cadena de suministros si los productos incluyen componentes de otros proveedores, también es necesario implementar y monitorizar métodos y procesos que validen que la información, y los productos se ajustan a los requisitos de seguridad establecidos y que identifiquen los componentes que sean críticos de mantener, ya que por lo tanto requerirán una mayor atención cuando se levanten los servicios. Además con estos procesos, la organización debe asegurarse de que los productos funcionan como se espera y no muestran ningún comportamiento anormal, esto incluye el manejo de riesgos de los componentes que dejen de estar disponibles a causa de proveedores con los que ya no se mantienen negocios, o

proveedores que ya no ofrecen ese componente por considerarlo obsoleto.

*15.2.1 Monitorización y revisión de los servicios de proveedores*

La monitorización y revisión de los proveedores de servicios debe asegurarse de que los términos y condiciones de seguridad de la información de los acuerdos se cumplen y que la información de los incidentes de seguridad se maneja adecuadamente. Esto implica que se deben supervisar los niveles de rendimiento del servicio, los informes proporcionados por el proveedor y organizar reuniones periódicas si fuera necesario para analizar estos informes de servicio, realizar seguimientos de las cuestiones que hayan sido detectadas en las auditorias, proporcionar información sobre los incidentes de seguridad, y gestionar cualquier problema identificado. Para poder realizar todas estas labores de una manera eficiente y organizada, la responsabilidad de gestionar las relaciones con los proveedores debería asignarse a un equipo de trabajo concreto.

*16.1.5 Respuesta a incidentes de seguridad de la información*

La respuesta a los incidentes de seguridad de la información debería incluir la recolección de evidencias tan pronto como sea posible, la realización de análisis forenses, el escalamiento si es necesario y la comunicación del incidente a todas las personas que se puedan ver involucradas y/o afectadas de alguna manera. Además la organización debe asegurarse de que todas las actividades de respuesta involucradas están debidamente registradas para su análisis posterior. Una vez que se ha tratado de manera satisfactoria el incidente, se debe grabar para posibles consultas en el futuro.

*17.1.1 Plan de continuidad de seguridad de la información*

La organización debe determinar los requisitos para la seguridad de la información y la continuidad de ésta en el tiempo, decidiendo si la continuidad de la seguridad de la información es capturada dentro del proceso de administración de continuidad del negocio o dentro del proceso de administración y recuperación de incidentes, los requisitos de seguridad de la información deben concretarse al planificar la continuidad del negocio y la recuperación de desastres, en ausencia de un plan de continuidad, la gestión de seguridad de la información debe asumir que los requisitos de seguridad de la información siguen siendo los mismos en situaciones adversas, en comparación con las condiciones operacionales normales.

*17.1.2 Implementación del plan de continuidad de seguridad de la información*

La organización debe establecer, documentar, implementar y mantener procesos, procedimientos y controles que aseguren el nivel de seguridad de la información en situaciones adversas. Para ello, la organización debe asegurarse de que dispone de personal con las competencias necesarias para mantener y administrar la seguridad de la información y de que se desarrollan planes y procedimientos de respuesta y recuperación de incidentes, que detallan como la organización manejará un evento adverso y mantendrá la seguridad de la información en un nivel predeterminado.

*17.1.3 Verificar, revisar y evaluar la continuidad de la seguridad de la información*

De manera frecuente, la organización debe verificar los controles de continuidad de seguridad de la información para asegurar que sean válidos y efectivos durante situaciones adversas.

*17.2.1 Disponibilidad de las instalaciones de procedimiento de información*

Con la finalidad de asegurar la disponibilidad de sistemas de la información, la organización debe identificar los requisitos de negocio, si no fuera posible garantizar la disponibilidad usando la arquitectura existente se debe considerar tener componentes o arquitecturas redundantes, éstos deben probarse para asegurar que la conmutación por fallos de un componente a otro funcione según lo previsto.

## 3.3. Anonimización

Anonimizar los datos personales mitiga los riesgos de seguridad, permite un tratamiento de la información más eficiente y seguro en entornos Big Data y refuerza la seguridad en Cloud Computing. La anonimización es el proceso irreversible de modificación y/o eliminación de datos personales de manera que no se pueda obtener información sujeta a las leyes de protección de datos, especialmente de carácter personal.

### 3.3.1. Agencia Española de Protección de Datos

La Agencia Española de Protección de Datos [2] establece los siguientes principios a tener en cuenta y determina que los actores que intervienen en el proceso de anonimización son:

- Responsable de fichero
- Destinatario de la información anonimizada
- Responsables técnicos del proceso de anonimización

Antes de anonimizar un conjunto de datos, se debe realizar una evaluación del impacto que tendría en la protección de datos personales y definir los objetivos que se quieren obtener, así como la finalidad de la información anonimizada. Esta evaluación es muy importante ya que cuando un conjunto de datos se anonimiza realmente y no es posible identificar a las personas, la legislación europea de protección de datos deja de ser aplicable en dicho conjunto de datos. Para ello es necesario tener en cuenta que:

- Se debe considerar el cumplimiento de los principios de protección de datos en el origen considerando el tipo de información recogida, la seguridad y los derechos ARCO. [1]
- Se debe hacer un análisis con el fin de identificar los riesgos teniendo en cuenta la naturaleza de los datos recogidos, el cruce con otras fuentes de información y la trazabilidad inversa.
- Garantizar la irreversibilidad del proceso en la medida de lo posible teniendo en cuenta la privacidad desde el diseño.

### 3.3.2. Relación de Técnicas de Anonimización

Se pueden clasificar las técnicas de anonimización en dos grandes familias: Aleatorización y Generalización. Estas técnicas se analizan siguiendo los criterios de riesgo que utiliza el "Grupo de Trabajo" sobre proteccion de datos del artículo 29, en el Dictamen 05/2014 sobre técnicas de anonimización [8].

Estos criterios son:

- **Singularización**: posibilidad de extraer registros que identifiquen a una persona de un conjunto de datos.
- **Vinculabilidad**: capacidad de vincular dos registros o más, de una o varias personas en la misma base de datos o en bases de datos diferentes.

[1] Acceso, Rectificación, Cancelación, Oposición

- **Inferencia**: posibilidad de deducir con una alta probabilidad el valor de un atributo a partir de los valores de otros atributos.

A continuación se describen las familias mencionadas.

#### 3.3.2.1. Técnicas de Anonimización de la Familia Aleatorización

Esta familia de técnicas se basa en la modificación de la veracidad de los datos, con el fin de eliminar el vínculo que les relaciona con las personas, de forma que si un dato es lo suficientemente ambiguo no se puede establecer una relación directa con una persona o grupo de personas en concreto. La aleatorización por si sola no reduce la singularidad pero si es eficaz contra los riesgos de inferencia.

La aleatorización se puede conseguir mediante 3 técnicas:

1. **Adición de Ruido**: modifica los atributos del conjunto de datos para que sean menos precisos, el nivel de ruido depende de la cantidad y tipo de información. Es un error pensar que la adición de ruido es una medida suficiente, ésta ha de verse mas bien como una medida complementaria que dificulta la acción de un atacante. Además si el ruido es inconsistente a nivel semántico el atacante podría filtrarlo fácilmente.

2. **Permutación**: mezcla los valores de los atributos para vincular algunos de ellos a distintas personas de manera artificial, intercambiando los valores contenidos en un conjunto de datos de un registro a otro. Es importante permutar un conjunto de atributos que estén relacionados entre sí, de otra forma un atacante podría identificar los atributos permutados y revertir los cambios, por este motivo es un error permutar los atributos de manera aleatoria. Esta técnica debe combinarse siempre con la eliminación de atributos clave y no dar por hecho que es una medida suficiente para anonimizar un conjunto de datos.

3. **Privacidad Diferencial**: a diferencia de las técnicas anteriores, la privacidad diferencial no modifica los datos originales. Suele usarse cuando el responsable del tratamiento genera vistas anonimizadas de un conjunto de datos y a la vez conserva una copia de los datos originales. Estas vistas anonimizadas se generan mediante un subconjunto de consultas a las que se les introduce ruido de manera aleatorio y son realizadas por un tercero. Los resultados de las consultas deben

considerarse datos personales en cuanto a que el responsable de tratamiento conserva los datos originales y por tanto pude identificar a las personas. Por ello es importante que se supervise de manera continua cualquier posibilidad de identificación de una persona en el conjunto de resultados de las consultas. Para reforzar la seguridad, el responsable de tratamiento puede y debe conservar una lista de todas las consultas realizadas, con el propósito de que los terceros no accedan a datos a los que no están autorizados, además a las consultas se les pueden aplicar técnicas de anonimización. Para reducir el riesgo de inferencia y vinculabilidad es necesario llevar un seguimiento de estas consultas y examinar la información que se ha obtenido, por lo que las bases de datos con privacidad diferencial no deberían utilizar motores de búsqueda que no permitan la trazabilidad. Es un error no añadir suficiente ruido para evitar la vinculación con conocimientos previos que se tengan sobre las personas o grupos. El mayor reto de esta técnica, consiste en generar la cantidad adecuada de ruido para añadir a las respuestas verdaderas, con la finalidad de proteger la privacidad de las personas, pero que al mismo tiempo preserve la utilidad de los datos. La Figura 3.1 muestra un ejemplo.

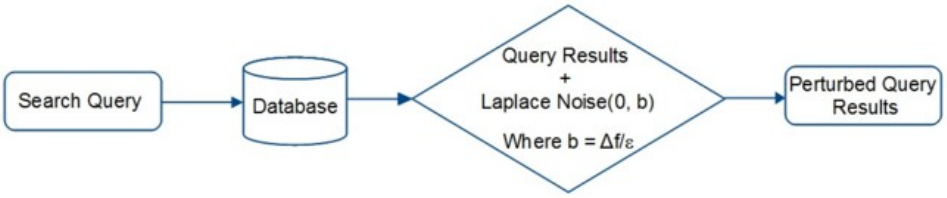


Figura 3.1: Como se aplica la privacidad diferencial
Fuente:http://ebusinesshoy.com/como-hacer-privacidad-diferencial-en-tu-ecommerce-igual-que-apple-google-y-microsoft/

#### 3.3.2.2. Técnicas de Anonimización de la Familia Generalización

Esta familia de técnicas se basa en la modificación de los atributos alterando las escalas u órdenes de magnitud. Las técnicas de Generalización son efectivas contra la singularización pero no garantizan una anonimización eficaz en todos los casos, como en el caso anterior es necesario combinar varias técnicas para prevenir también la vinculabilidad y la inferencia.

La aleatorización se puede conseguir mediante 3 técnicas:

*a*) **K-Anonimato y Agregación**: estas técnicas tienen como objetivo que una persona pueda ser singularizada cuando se la agrupa al menos con k personas más, para ello se generalizan los valores de los atributos hasta que todos acaban compartiendo el mismo valor. El punto crítico de esta técnica consiste en saber determinar el valor de K, cuanto más alto sea más garantías de privacidad obtendremos, pero si es demasiado elevado se reduce el número de identificadores, lo que provoca que algunas personas no estén protegidas por la generalización. Por otro lado un valor demasiado bajo de K hace que el peso de cualquier persona aumente por lo que el riesgo de inferencia aumenta.

*b*) **Diversidad-l y Proximidad-t**: esta técnica extiende al anonimato K para evitar los riesgos de inferencia (siempre que los valores de los atributos estén bien distribuidos), para lograrlo todos los atributos tienen que tener al menos L valores diferentes. La proximidad-t perfecciona a la diversidad-L, en este caso ya no vale con que haya L valores diferentes, sino que además cada valor debe representarse tantas veces como sea necesario para que se refleje la distribución inicial de cada atributo.

### 3.3.3. La pseudonimización y sus riesgos

Uno de los errores más frecuentes consiste en pensar que la pseudonimización es equivalente a la anonimización, esta técnica consiste en reemplazar un atributo por otro en un registro, lo que sigue permitiendo identificar a una persona. Si es cierto que la pseudonimización reduce la vinculabilidad de un conjunto de datos, pero debe entenderse como una medida de seguridad y no como un método de anonimización.

Por otro lado está la cuestión de confianza en el controlador de los datos, que asume toda la responsabilidad legal en cuanto a la liberación de los datos y las técnicas de anonimización elegidas. Si se elige este enfoque centralizado de anonimización los interesados no pueden verificar si su información obtiene una protección adecuada, mientras que si se opta por una anonimización local, en la que el responsable de anonimizar los datos es su propietario, entramos en la dicotomía sobre el nivel de anonimización y privacidad adecuado, que por lo general tenderá a ser muy elevado, reduciendo así el valor de los datos. En la

Tabla 3.1 se muestra un pequeño resumen comparativo de las técnicas de anonimización y la pseudonimización.

#### 3.3.3.1. Técnicas de Pseudonimización

- Cifrado con clave secreta: En esta técnica el poseedor de la clave es capaz de identificar a la persona descifrando el conjunto de datos, si se aplicaran los sistemas de cifrado más avanzados, sólo se podrían descifrar los datos si se conoce la clave.

- Función Hash: Este tipo de funciones devuelven un resultado de tamaño fijo a partir de un valor de entrada de cualquier tamaño, aunque estas funciones no son reversibles, es decir, no existe el riesgo de revertir el resultado como pasa con el cifrado, son vulnerables a los ataques por fuerza bruta, si se conoce el rango de los valores de entrada, se pueden pasar estos valores por la función con el objetivo de obtener el valor real de un registro determinado.

  Función Hash con sal: se añade un valor aleatorio al atributo al que se aplica la función hash para reducir la probabilidad de obtener el valor de entrada.

- Función con clave almacenada: Se trata de un tipo de función hash que hace uso de una clave secreta como si fuera un valor de entrada suplementario, en este caso el responsable del tratamiento puede reproducir la ejecución de la función con el atributo y la clave secreta.

- Cifrado Determinista o Función Hash con borrado de clave: Esta técnica equivale a generar un número aleatorio para cada atributo de la bases de datos como si fuera un seudónimo para poder borrar la tabla correspondiente, con esta técnica se reduce el riesgo de vinculabilidad entre los datos personales contenidos e otro conjunto de datos en que se usará un seudónimo diferente.

- Descomposición en Tokens: Esta técnica usada en el sector financiero reemplaza los números de identificación de tarjetas por valores de poca utilidad para el atacante, y se basa en la aplicación de mecanismos de cifrado unidireccionales y en la asignación mediante una función de índice, de un número de secuencia o de un número generado de manera aleatoria para que no derive matemáticamente en los datos originales.

Tabla 3.1: Resumen comparativo de las técnicas de anonimización con la pseudonimización.

| | *Singularización* | *Vinculabilidad* | *Inferencia* |
|---|---|---|---|
| *Adición de Ruido* | si | puede que no | puede que no |
| *Permutación* | si | si | puede que no |
| *Privacidad Diferencial* | puede que no | puede que no | puede que no |
| *k-Anonimato y Agregación* | no | si | si |
| *Diversidad-L y Proximidad-T* | no | si | puede que no |
| *Pseudonimización* | si | si | si |

# Capítulo 4

# Especificaciones para auditorías de proyectos con tecnologías Open Data

## 4.1. Introducción

Open Data es un movimiento internacional, que pretende que los datos de las administraciones públicas (AAPP) estén disponibles para todos los ciudadanos, para que se pueda acceder libremente a ellos y reutilizarlos generando así nuevos productos útiles para el conjunto de la ciudadanía. El concepto relativamente nuevo de datos abiertos tiene beneficios para la sociedad, ya que promueve la democracia de los datos fomentando la transparencia de gobiernos y organizaciones. Estos grandes beneficios son motivo suficiente para desarrollar estándares y principios de buenas prácticas que se deben tener en cuenta con el fin de garantizar la calidad, accesibilidad, y disponibilidad de los datos. No obstante, la adopción de este movimiento supone la adaptación a un cambio de paradigma.

## 4.2. Características del Open Data

A continuación se listan las características que la información debe poseer para que pueda considerarse abierta. Esta información está publicada en el Trabajo Fin de Máster: *Auditoría y propuesta de meto-*

*dología para publicación de datos abiertos en ciudades inteligentes* [12].

- **Disponibilidad y Accesibilidad**: La información debe estar accesible en internet, y debe estar disponible la mayor cantidad de tiempo posible.
- **Reutilización y Redistribución**: Con la intención de que la información sea reutilizable, sus condiciones de uso deben permitir que ésta se pueda modificar y que el resultado de la modificación se pueda distribuir libremente. Además se deben facilitar formatos digitales lo más estandarizados y reutilizables posible.
- **Universalidad**: Los datos deben estar disponibles para cualquier persona que desee utilizarlos independientemente del fin con el que quiera hacerlo.

### 4.2.1. Open Knowledge Foundation

Esta fundación sin ánimo de lucro promotora del conocimiento abierto, sostiene que los principios del Open Data deben ser los siguientes:

- Acceso libre y gratuito.
- Libertad de distribución.
- Libertad de reutilización.
- Ninguna restricción sobre los puntos anteriores basándose en la persona, lugar o sector empresarial.

### 4.2.2. Sunlight Foundation

La Sunlight Foundation es una entidad sin ánimo de lucro, fundada en 2006 y con sede en Washington D.C cuyo objetivo es fomentar una mayor apertura y transparencia del gobierno en Estados Unidos. [25] Según esta organización, el Open Data debe cumplir con los siguientes principios:

- Completitud: Todos los datos que no estén sometidos a limitaciones de protección de la privacidad, seguridad y/u otras características deben de estar disponibles.
- Datos primarios: Los datos publicados deben tener el mayor nivel de granularidad posible.
- Oportunidad y puntualidad: Se deben publicar los datos tan pronto como sea posible para preservar su valor.

- Facilidad de acceso: Los datos deben estar disponibles para el mayor número posible de personas, esto incluye registros tanto físicos como digitales.
- Procesable informáticamente: Los datos deben estar estructurados de tal manera que permitan el procesamiento automático.
- No discriminación y accesibilidad: Los datos deben estar disponibles para todo el mundo sin importar condición alguna, ni siquiera la necesidad de registrarse.
- Estándares abiertos: Los datos deben estar en formatos no propietarios, o de uso común gratuitos.
- Licenciamiento gratuito: Los datos no deben estar sujetos a ningún tipo de copyright o patente, aunque se pueden imponer condiciones de seguridad y privilegios.
- Permanencia: Los datos publicados deben permanecer en línea con los controles adecuados de versiones.
- Coste del uso: Los datos publicados no deben suponer un coste alguno para el usuario.

## 4.3. Cadena de Valor del Open Data

Siguiendo el criterio de Alberto Abellá en *Reutilización de información pública y privada en España* [1], la cadena de valor del Open Data como se muestra en la Figura 4.1 está dividida en cinco etapas.

*a*) **Fuentes de Datos o publicadores:** Organismos públicos o privados productoras de los datos.

*b*) **Esquemas legales y técnicos:** Esquemas que permitan la publicación de los datos, legales en cuanto a la reutilización y técnicos en cuanto al cómo se publican y se reutilizan.

*c*) **Infomediarios:** Entidades que generan nuevos servicios a partir de los datos ya publicados, estas entidades pueden ser empresas, ciudadanos o entidades sociales.

*d*) **Productos y Servicios:** Aplicaciones, informes, nuevos servicios de carácter económico o social.

*e*) **Usuarios:** Usuario profesional, ciudadano anónimo.

## 4.4. Métricas

Ante la falta de homogeneidad actual en los conjuntos de datos publicados en los portales Open Data, las métricas que nos ayudan a


Figura 4.1: Cadena de valor del Open Data
Fuente: https://datosenabierto.wordpress.com/ 2017

conseguir que los datos se publiquen de manera apropiada son diversas y variadas. A continuación analizamos algunas de ellas.

### 4.4.1. Cinco Estrellas

Tim Berners-Lee analiza los conjuntos de datos en base al formato y a la licencia con la que han sido publicados, según esta métrica [13] los datos deben estar integrados en la web además de enlazados con otros datos, como su nombre indica contempla cinco niveles siendo una estrella el nivel más básico y cinco estrellas el más alto. El criterio subyacente de esta métrica es que la información sea accesible a través de la web bajo una licencia abierta para que el usuario pueda visualizarla, guardarla, utilizarla y reutilizarla libremente. La Figura 4.2 muestra cada uno de estos niveles.

- Una Estrella: Datos publicados con formatos no estructurados con una licencia abierta como PDDL, ODC-by o CCO pero de los que es difícil extraer información, como por ejemplo un pdf o jpg.
- Dos Estrellas: Datos accesibles desde la web en formato estructurado pero para extraer información es necesario un software propietario, como pasa con los .doc
- Tres Estrellas: Datos publicados usando formatos estructurados y abiertos que cualquiera puede utilizar y manipular, sin embargo no están integrados en la web; como puede ser un XML.
- Cuatro Estrellas: Los datos están integrados en la web y se pueden identificar mediante una URI, pudiendo hacer referencia a los mismos desde otras bases de datos.

- Cinco Estrellas: Los datos están integrados en la web y enlazados a otros datos, se puede hacer referencia a datos de la misma o de distinta entidad usando la URL, pero el formato debe ser legible mediante el modelo RDF.

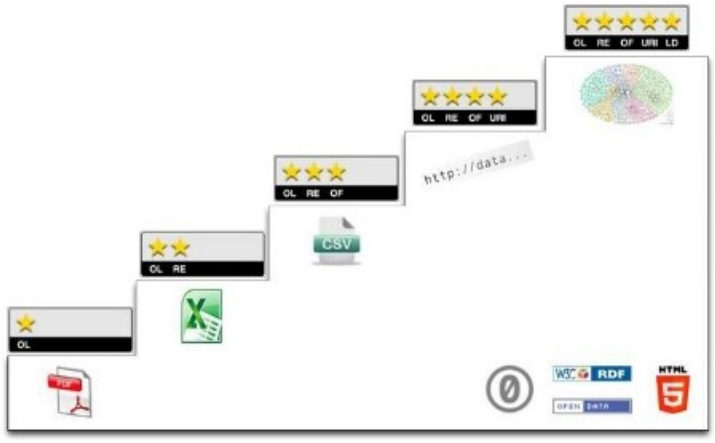


Figura 4.2: Cinco Estrellas del Open Data, Tim Berners-Lee
Fuente: http://5stardata.info/en/ 2017

### 4.4.2. Meloda

Esta métrica [11] creada por Alberto Abellá permite evaluar el grado de reutilización de los datos abiertos, está compuesta por seis dimensiones (Estándares técnicos, Acceso a la Información, Marco legal, Modelos de datos, Geolocalización de la información y Actualización en tiempo real), cada dimensión tiene niveles que se asignan según el cumplimiento del conjunto de datos y cada nivel tiene asociado un peso, que es un valor porcentual que se utiliza para calcular el valor Meloda del conjunto de datos.

- **Estándares técnicos:** En esta dimensión se evalúan los formatos en los que se publican los conjuntos de datos. Se califica la información en 4 niveles y se les asigna un peso a cada uno de ellos:
  - Nivel 1: El conjunto de datos está publicado en un formato estándar cerrado no reutilizable (peso = 10 %) Ej: pdf, doc.
  - Nivel 2: El conjunto de datos está publicado en un formato estándar cerrado reutilizable (peso=35 %) Ej: xls.

- Nivel 3: El conjunto de datos está publicado en un formato estándar abierto (peso= 60 %) Ej: txt,odt, ods.
- Nivel 4: El conjunto de datos está publicado en un formato estándar abierto con metadatos (peso= 100 %) Ej: rdf, rss, json, xml.

- **Acceso a la Información:** En esta dimensión se evalúa la facilidad con la que se accede a la información.
  - Nivel 1: Si el acceso a la información no es libre, si no hay acceso web, se requiere una solicitud manual para poder acceder a la información o los datos están registrados en un formato no digital (peso=0 %).
  - Nivel 2: Acceso vía web con registro, el acceso a la información es a través de la web, pero requiere un registro de usuario para poder seleccionar los datos (peso=10 %).
  - Nivel 3: Acceso directo vía web con una URL única y constante, la URL tiene parámetros (pero no variación en los parámetros) para el mismo conjunto de datos (peso=50 %).
  - Nivel 4: Acceso vía web con parámetros, la URL tiene parámetros y no es necesario tener que descargar todo el conjunto de datos (peso=90 %).
  - Nivel 5: Acceso completo, se ofrece una API o lenguaje de consulta (peso=100 %).
- **Marco legal:** En esta dimensión se mide el tipo de licencia con el que se han publicado los conjuntos de datos.
  - Nivel 1: Copyright: el autor se reserva el derecho de los datos, restringe el uso no autorizado (peso=0 %).
  - Nivel 2: Uso privado: se permite el uso de datos sin necesidad de aprobación, pero solo para usos privados (peso=10 %).
  - Nivel 3: Reutilización no comercial: se permite la reutilización de datos, siempre y cuando no esté destinada a un uso comercial (peso=25 %).
    - (CC BY-NC-ND 4.0) Creative Commons Attribution - NonCommercial - NoDerivatives 4.0
    - (CC BY-NC 4.0) Attribution-NonCommercial 4.0 International
    - (CC BY-NC-SA 4.0) Attribution-NonCommercial-ShareAlike 4.0 International
  - Nivel 4: Reutilización comercial: se permite la reutilización de datos, incluidos los usos comerciales (peso=90 %).
    - (CC BY-ND 4.0) Attribution-NoDerivatives 4.0 International

    - (CC BY-SA 4.0) Attribution-ShareAlike 4.0 International
  - Nivel 5: Sin restricciones: sólo se pide la atribución de los datos a los reutilizadores (peso=100 %).
    - Attribution 4.0 International (CC BY 4.0)

- **Modelos de datos:** Esta dimensión analiza la capacidad para compartir su modelo de datos con otras entidades.
  - Nivel 1: Sin modelo publicado: El modelo no se libera ni se explica (peso=15 %).
  - Nivel 2: Modelo con campos de datos: Hay campos diseñados por el editor, éstos se identifican simplemente pero no se explican los rangos, tipo y resto de características (peso=35 %).
  - Nivel 3: Modelo de datos propio con especificaciones de campos: Se detallan y se publican las especificaciones de los campos, aunque sea propio; y se deja libre para ser utilizado por otros. Ej: Hay una descripción pública disponible así como vocabularios (peso=50 %).
  - Nivel 4: Modelo local de datos abiertos: Se usa un modelo externo normalizado, aunque esté poco extendido (peso=90 %).
  - Nivel 5: Modelo de datos abiertos: Hay un modelo de datos normalizado publicado por una entidad internacional como puede ser ISO y/o existe una amplia adopción (peso=100 %).

- **Geolocalización de la información:** La información publicada podría contener referencias sobre su localización. Esto no implica que la información tenga que estar geolocalizada, sino que tiene algunos campos que permiten identificar la ubicación de la información.
  - Nivel 1: No hay información geográfica: La información publicada no tiene ningún campo que haga referencia a su ubicación (peso=15 %).
  - Nivel 2: Campo de texto simple: La información geográfica es sólo un campo de texto o un Id propio que hace difícil conectar esa información con otras bases de datos (peso=30 %). Ej: El campo País, valor España.
  - Nivel 3: Campo de texto complejo: La información geográfica está compuesta por varios campos con descripción de texto que además son jerárquicos (peso=50 %).
  - Nivel 4: Coordenadas: La información publicada incluye dos campos con las coordenadas (peso=90 %).
  - Nivel 5: Nivel de información geográfica completa: Incluye los niveles 3 y 4 (peso=100 %).

- **Actualización en tiempo real:** Esta dimensión mide la frecuencia de actualización de los conjuntos de datos, que puede variar en función de su naturaleza.
  - Nivel 1: Semanas: Cuando el periodo de actualización del conjunto de datos es mayor que una semana (peso=15 %).
  - Nivel 2: Días: Cuando el periodo de actualización oscila entre uno y siete días (peso=40 %).
  - Nivel 3: Horas: Cuando el periodo de actualización oscila entre una y veinticuatro horas (peso=70 %).
  - Nivel 4: Minutos: Cuando el periodo de actualización oscila entre un minuto y una hora (peso=90 %).
  - Nivel 5: Segundos: El periodo de actualización es inferior a un minuto (peso=100 %).

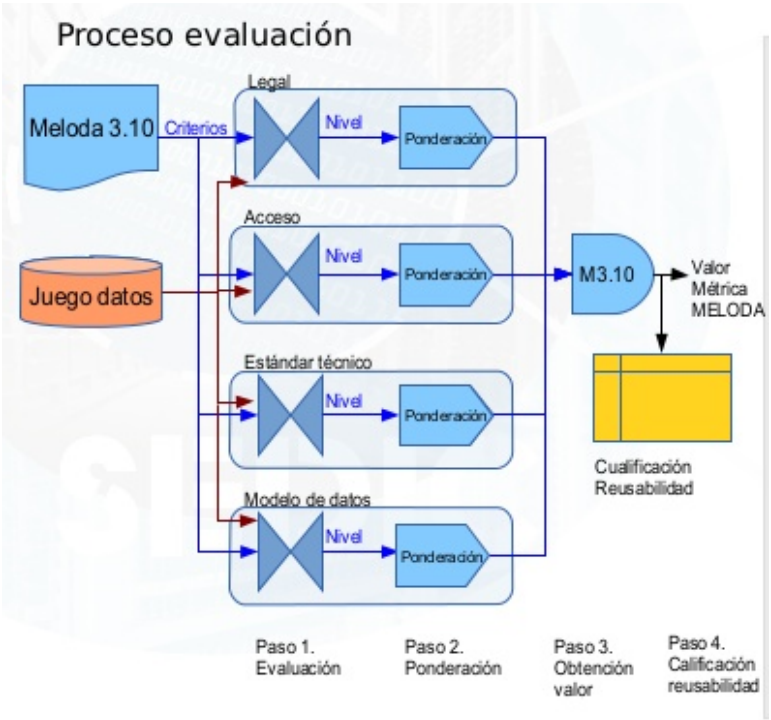


Figura 4.3: Procesos de Meloda, Alberto Abellá
Fuente: http://www.meloda.org 2017

La Figura 4.3 muestra como aplicar esta métrica a un conjunto de datos, en particular hay que seguir los siguientes pasos:

*a*) Analizar el conjunto de datos desde el punto de vista de cada una de las dimensiones explicadas anteriormente y asignarles el nivel

Tabla 4.1: Rangos de clasificación de Meloda

| Rangos | Clasificación |
|---|---|
| 0-25 | Inadecuado para la reutilización |
| 25-50 | Posible reutilización básica |
| 50-75 | Reutilización avanzada pero con características mejorables |
| 75-100 | Reutilización avanzada |

que les corresponde.

*b*) Asignar el peso correspondiente, según el nivel asignado en el paso anterior.

*c*) Calcular el valor Meloda como cien veces la raíz cúbica del producto de los pesos de cada dimensión.

*d*) El resultado obtenido de aplicar la fórmula anterior será un valor numérico comprendido entre 0 y 100, cuyo valor se lo compara en la columna de rangos de la Tabla 4.1 y se obtiene la calificación de reutilización para el conjunto de datos.

### 4.4.3. Norma española: Ciudades Inteligentes. Datos Abiertos

La norma UNE 178301: 2015 [15] permite evaluar la publicación de datos abiertos en organismos del sector público en el ámbito de las ciudades inteligentes, la norma se estructura en cinco dominios que a su vez se dividen en dimensiones, (que en este resumen se han obviado para facilitar la comprensión) dentro de las cuales se encuentran las métricas que a continuación se muestran.

*a*) **Dominio Estratégico**

Este dominio establece los criterios para evaluar la capacidad y ejecución de la política estratégica del organismo para articular una visión consistente de la apertura de datos.

- Estrategia: El organismo debe establecer una estrategia de apertura de datos a nivel político, y se debe documentar el plan estratégico para la apertura de datos incluyendo al propio organismo y a organismos dependientes. La estrategia debe estar alineada con las estrategias de datos abiertos de otros niveles de la administración (comunidad autónoma, nacional, europeo) y debe implantarse y medir el cumplimientos del

proceso. La información a contemplar que debe establecer la estrategia a rasgos generales consiste en la exposición general de los conjuntos de datos, la priorización de publicación de datos, condiciones de reutilización (generales y especificas), medición del cumplimiento, apertura de datos como una fase adicional del tratamiento de los sistemas de información del organismo, gestión del cambio que implica este proyecto, colaboración y participación con reutilizadores.

- Liderazgo: El organismo debe disponer de suficiente capacidad de liderazgo para conducir con éxito un proceso de apertura de datos, lo que significa que el organismo debe, asignar las funciones de apertura de datos a un responsable político y dotarse de capacidad de actuación suficiente sobre el resto de departamentos del organismo.

- Compromiso de servicio: El organismo debe crear una carta de servicios que tendrá que estar documentada, ser de carácter público y estar publicada para la apertura de datos, y que contendrá a grandes rasgos los derechos que asisten a los ciudadanos, los compromisos de calidad y la responsabilidad de los gestores públicos ante la ciudadanía.

- Sostenibilidad Económica : El organismo debe asegurarse de que el proyecto es sostenible económicamente, para ello la evaluación de la sostenibilidad económica debe documentarse en un estudio donde la información a contemplar debe incluir la existencia de recursos económicos, técnicos y humanos, el coste de las aplicaciones, producción y gestión de datos, la posible existencia de un modelo de cobro por los datos y todos aquellos instrumentos que se consideren necesarios para asegurar la sostenibilidad económica. Además dicha sostenibilidad debe contemplar las fases de puesta en marcha (estudio a corto plazo, un año máximo), mantenimiento y evolución.

*b*) **Dominio Legal:**
Este dominio establece los criterios para evaluar en el organismo la existencia y verificación de leyes y normativas que facilitan la ejecución de las políticas y actividades. Además establece los criterios para evaluar la capacidad del organismo de cumplir con la legislación vigente en materia de apertura de datos, de elaborar la propia si fuera necesario y de establecer las condiciones legales del uso de los datos.

- Normas externas e internas: El organismo debe cumplir la normativa referente a datos abiertos, es decir cumplir la normativa vigente de orden superior obligatoria (Ley 37/2007 y la Norma Técnica de Interoperabilidad de Reutilización de la Información del Sector Publico). Adicionalmente, debería cumplir la normativa vigente de orden superior recomendada (Real Decreto 1495/2001) y seguir las normas y guías que permitan una armonización entre todas las Administraciones que compartan los mismos principios y definiciones para asegurar la interoperabilidad y aprovechamiento eficiente de las sinergias llevadas a cabo por todos los actores. Además sería recomendable que el organismo creara y publicara normativas propias que maticen o concreten la normativa de orden superior.

- Condiciones de uso y Licenciamiento: El organismo debe establecer las licencias y condiciones de uso de los datos para la reutilización, estas licencias y condiciones deben ser conformes con la legislación vigente (Real Decreto 1495/2011), estar documentadas y publicadas en la web, ser claras, transparentes y no discriminatorias, además de permitir un uso abierto, libre y sin restricciones de uso comercial. Por ello se deberían usar licencias estándar, licencias auto documentadas y preferiblemente procesables electrónicamente.

*c*) **Dominio Organizativo:**
Este dominio comprende actividades que permiten determinar si el organismo cuenta con el equipo de trabajo adecuado que lleve a la práctica la iniciativa de los datos abiertos.

- Unidad Responsable: El organismo debe tener una unidad responsable de la apertura de datos, que estará definida en el organigrama o similar y que tiene como objetivos el desarrollo, evolución y seguimiento de la normativa, la elaboración de circulares y recomendaciones y la coordinación con las áreas organizativas. Además entre sus funciones se encuentran tareas como las derivadas de la ejecución de la apertura de datos, recibir peticiones de datos, analizar la disponibilidad de los mismos y formalizar su distribución.

- Equipo de trabajo y capacitación: La unidad responsable de datos abiertos debe disponer de un equipo de trabajo que

tiene la responsabilidad de la implantación y ejecución del proyecto de los datos abiertos en organismo. Ejecuta las tareas de gestión y planificación, identificando la información publicable, priorizando la información, creando documentación. Además el equipo de trabajo debería estar adecuadamente dimensionado en función del tamaño y/o complejidad del organismo, debería tener formación específica en normativa y técnicas de datos abiertos así como seguir lo especificado en las Guías de aplicación de la legislación vigente (Guía de aplicación del Real Decreto 1495/2011).

- Inventario: La unidad responsable de datos abiertos debe generar un inventario de información reutilizable. El inventario debe incluir la relación y descripción de las bases de datos utilizadas en los sistemas de información del organismo, así como la relación y descripción de las bases de datos generadas por actividades auxiliares relacionadas (estadísticas, investigación). Como información adicional, se debería incluir la posibilidad de ser publicado, la disponibilidad técnica para la publicación, el periodo de actualización, la unidad organizativa responsable y contacto para sugerencias y quejas de la fuente. En cualquier caso, el inventario debe documentarse.

- Prioridad: La unidad responsable debe establecer la prioridad para la publicación de los datos recogidos en el inventario. El método de priorización debe documentarse, y la prioridad debe estar definida en base a criterios de relevancia, demanda y calidad.

- Medición de cumplimiento del proceso: La unidad responsable de datos abiertos debe tener un plan de evaluación que mida el nivel de cumplimiento del plan estratégico de apertura de datos. La medición debe documentarse y realizarse de manera periódica, la información a contemplar debe incluir aunque no de manera exhaustiva el trabajo realizado en cuanto a la evolución del proyecto, gestión del cambio, el nivel de coordinación con otros departamentos, las dificultades organizativas o técnicas que se hayan encontrado para disponer de los datos, y las propuestas de mejora como pueden ser la optimización de la gestión, los cambios en los criterios de prioridades además de las necesidades de promoción.

- Medición del uso e impacto: La unidad responsable de da-

tos abiertos debe medir el nivel de acceso y uso que se está haciendo de los datos publicados, esta medición debe documentarse y realizarse de manera periódica. La información a contemplar debe incluir la evolución del número de descargas total, por categoría y por formato, la evolución del número de conjuntos de datos disponible total por categoría, y formato, la evolución del número de peticiones de conjuntos de datos total, por categoría y formato, la evolución del número de aplicaciones desarrolladas internas o externas por conjuntos de datos utilizados y el tipo de aplicación.

*d*) **Dominio Técnico:**
Este dominio establece los criterios para evaluar aquellas actividades relacionadas para garantizar la existencia de protocolos y mecanismos que garanticen la disponibilidad de los datos en cada momento, y las actividades relacionadas para mantener y gestionar la calidad de los datos y el grado de interoperabilidad.

- Catálogo: El organismo debe publicar un catálogo de datos propio que proporcione el acceso a los conjuntos de datos mediante una interfaz web e interfaz de consulta accesible por HTML, y procesable por máquinas según la legislación vigente (Norma Técnica de interoperabilidad de Reutilización de recursos de la Información).

- Presencia en el Catálogo de Información Pública: El organismo debe realizar las acciones precisas para publicar los datos en el Catálogo de Información Pública, (www.datos.gob.es).

- Conjuntos de datos documentados: Los conjuntos de datos publicados en el catálogo deben estar documentados mediante metadatos de acuerdo a la legislación vigente.

- Categorización y búsqueda: El catálogo debe disponer de mecanismos de búsqueda de los conjuntos de datos con criterios tales como texto, taxonomías, formato, categorías tipo de información, periodo de actualización, fecha de actualización, posibilidad de multicriterio.

- Disponibilidad: Se debe asegurar que los datos están disponibles mediante mecanismos de revisión de los enlaces en el catálogo y que el servicio sea de alta disponibilidad.

- Referencias persistentes y amigables: El catálogo debe estar configurado de forma que se asegure que los conjuntos de datos tienen una dirección permanente en el tiempo y fácilmente accesible. La dirección de un conjunto de datos debe ser accesible por HTTP, estable, indefinida y extensible. Y la estructura de la dirección de un conjunto de datos debe ser estable, homogénea, comprensible y estar definida de acuerdo a la legislación vigente (Norma Técnica de interoperabilidad de Reutilización de recursos de información).

- Accesibilidad/No discriminación: El organismo debe garantizar el acceso no discriminatorio a los datos, con medidas para garantizar los derechos de igualdad de oportunidades, no discriminación y accesibilidad según la legislación vigente. Para ello se debería admitir el registro de usuario en caso de que por necesidades de gestión o técnicas se la única forma de asegurar el servicio de distribución, en este caso se debe justificar la elección del sistema, la forma de gestionar el acceso, la posible existencia de controles o limitaciones y publicar en la web.

- Gratuidad: El organismo debe garantizar el acceso al dato de forma gratuita, para ello se debería admitir el pago solamente en caso de que por necesidades de gestión o técnicas sea la única forma de asegurar el servicio de distribución. Si esto ocurriera se debería justificar la elección de este sistema, así como la cuantía de las contraprestaciones económicas y el hecho de que la cuantía es marginal y orientada exclusivamente al coste de distribución, sin que los ingresos sirvan para cubrir los costes de recogida y producción y publicarlo en la web.

- Sistemas de acceso: El catálogo debe facilitar los sistemas de acceso a los datos de los conjuntos de datos adecuados para su uso, con medidas para permitir como sistema de acceso la descarga en forma de ficheros, una API web o un servicio SPARQL.

- Datos primarios: Los conjuntos de datos deben asegurar que los datos son primarios, es decir que tienen el mayor nivel de desglose posible, que los datos están recolectándose de la fuente origen y publicar los datos sin hacer tratamientos de agregación, modificación o resumen. Además los datos que incluyen datos protegidos de carácter personal o de seguridad

según la legislación vigente, deben ser tratados, pero de forma mínima para dar el máximo de desglose posible. Siempre que se publiquen los datos primarios, y como complemento, se pueden publicar datos agregados para facilitar otros usos.

- Datos completos: Los conjuntos de datos para que se consideren completos deben reflejar la totalidad del tema, estar descritos con el máximo detalle, estar cumplimentados todos los valores de los datos que pueden estarlo y respetar la protección de datos de carácter personal o de seguridad contemplada en la normativa vigente.

- Datos documentados: En el catálogo de datos debe publicarse la documentación de los datos, la forma de documentación debe ser en formatos en los que por su naturaleza solo se incluyen datos pero no la estructura y tipo, se debe documentar y publicar como complementario. En formatos auto documentados, donde la documentación de la estructura y contenido viene implícita y no es necesaria una documentación adicional, la información a contemplar debe incluir el significado del dato, la estructura, la referencia al dato completo si se trata de abreviaturas.

- Datos técnicamente correctos: El organismo debe realizar validaciones que aseguren que los datos son técnicamente correctos, estas validaciones deben documentarse y realizarse cuando se incorporen nuevos datos o se actualicen. La validación debería incluir si la sintaxis es correcta, si se utiliza la misma codificación y normalización para el mismo tipo de dato publicado en diferentes conjuntos de datos del catálogo y si la codificación y normalización utilizada se basa en algún estándar común reconocido y utilizado por otras organizaciones.

- Datos georreferenciados: El organismo debe facilitar que los datos geográficos publicados en el catálogo de datos estén asociados a una georreferenciación basada en coordenadas. Se considerará que existe una asociación explícita cuando están georreferenciados directamente mediante coordenadas en un sistema de referencia de coordenadas, y de manera implícita cuando se emplea una georreferenciación indirecta basada en un sistema de referencia basado en datos geográficos bien definidos.

- Datos enlazados: El organismo debe facilitar que los datos publicados estén enlazados con datos de otras fuentes de datos mediante enlaces RDF.

- Proceso de actualización: El organismo debe asegurar que el proceso de actualización sea el óptimo, este proceso de actualización debe contemplar la minimización de circunstancias que pueden afectar a la ejecución del proceso, verificar que se ha producido la actualización de forma correcta si el proceso es automático. Además este proceso debe documentarse.

- Frecuencia de actualización: El organismo debe asegurar que la frecuencia de actualización sea óptima, dicha frecuencia de actualización debe documentarse y debe tender a minimizar el tiempo que transcurre entre la actualización del dato en el sistema de origen y la actualización como datos abiertos. Preferentemente ambas deben tener una frecuencia similar.

- Ampliación de conjuntos de datos publicados: El organismo debe ampliar el conjunto de datos publicados, el plan de ampliación debe documentarse en forma de plan evolutivo y debe contemplar la causa, la priorización del organismo, la solicitud de publicación por parte de ciudadanos o reutilizadores y el establecimiento de una posible fecha de publicación, además este plan debe ser público y estar publicado en la web.

*e*) **Dominio Económico y Social:**

Este dominio, establece los criterios para evaluar los mecanismos que relacionan a los organismos que producen los datos con los reutilizadores, la compartición de estructuras comunes que fomenten la aplicación de los datos en la producción de nuevos servicios, el grado de implicación del organismo en el estímulo y ayuda a la labor de los agentes reutilizadores, el grado de escucha y adaptación a las demandas y el nivel de diálogo establecido.

- Cantidad de datos publicados: La organización debe publicar los conjuntos de datos identificados en el inventario, asignando a cada elemento uno de los elementos del inventario un solo conjunto de datos. Adicionalmente, se podrán publicar subconjuntos de datos para facilitar el acceso a elementos más concretos sin que en ningún caso se llegue a conjuntos de da-

tos con un solo dato.

- Formato de los datos: Los conjuntos de datos deben estar publicados en formatos que cumplan los estándares, que sean abiertos (no propietarios), que estén estructurados y que permitan la identificación única de los recursos. Hay que señalar que se permite la publicación en otros formatos propietarios siempre y cuando éste no sea el único formato utilizado. En igualdad de condiciones se priorizará aquellos formatos más adecuados para el uso por reutilizadores y con el menor coste de tratamiento e integración.

- Vocabularios: Los conjuntos de datos deben estar publicados utilizando vocabularios que estén expuestos públicamente, que tengan una URL persistente, que sigan convenios, esquemas y que sean auto descriptivos. Por otro lado, los vocabularios deberían tener una política de control de versiones y estar descritos en más de un idioma, además deberían estar publicados utilizando vocabularios estándares, es decir que se haya definido como estándares por una organización internacional, europea o nacional de normalización, que se haya definido como de uso prioritario en normativas y guías y que sea independiente del productor de los datos. Si los vocabularios utilizados son dependientes, (es decir no estándares) deben exponerse públicamente.

- Transparencia, participación y colaboración: El organismo debe facilitar la transparencia, participación y colaboración disponiendo de un canal específico para enviar opiniones, o pedir la publicación de nuevos datos o la mejora de los actuales. Y debería publicar los comentarios, permitir añadir conversaciones y publicar el resultado del análisis realizado por el organismo con una planificación para su implementación.

- Resolución de quejas y conflictos: El organismo debe facilitar la resolución de quejas y conflictos sobre datos abiertos cumpliendo con los parámetros de calidad del servicio de quejas y sugerencias que ya disponga.

- Fomento de la reutilización: El organismo debe fomentar la reutilización de los datos, mediante la educación en el uso de los datos tanto internamente como a los colectivos de reutilización, o el fomento del conocimiento y la inquietud por

procesar información de una forma autónoma. Para ello se deberían facilitar recursos de ayuda, documentos, materiales editados o promovidos por la organización y actividades formativas.

- Iniciativas de reutilización desarrolladas: El organismo debe desarrollar iniciativas que fomenten la reutilización de datos. En concreto favoreciendo acciones de desarrollo como concursos, premios e incorporación a las nuevas aplicaciones, desarrollos que se realicen en el organismo o la utilización datos abiertos que ya estén publicados.

### 4.4.4. Ojo al Data

Este proyecto tiene como objetivo determinar los cien conjuntos de datos más relevantes que deben ser tenidos en cuenta en una iniciativa Open Data para una smart city, tomando como referencia la ciudad de Madrid. El grupo de trabajo de Datos Abiertos de la Federación Española de Municipios y Provincias que cuenta con un total de diecisiete personas de distintos ámbitos entre las que se incluyen la directora de este proyecto la Dra. Victoria López López y la profesora de esta facultad Guadalupe Miñana, ha conseguido crear un modelo de referencia de conjuntos de datos común a todas las AAPP para facilitar la reutilización de la información. Donde se ha tenido en cuenta:

- Qué datos deben abrirse y cómo, según la tipología y medios de los municipios.
- Normalización de la estructura así como de los contenidos de los datasets.
- Interoperabilidad de los datos entre AAPP y organismos públicos.
- Datos Abiertos hacia fuera y hacia dentro de las organizaciones.
- Herramientas TIC para la puesta en marcha.
- Compartir portales de datos abiertos entre municipios y federación en portal nacional y europeo.
- Aplicación de normativa y creación de instrumentos legales.

También han definido la estrategia a seguir en Datos Abiertos para las AAPP, se ha elaborado una guía para su puesta en marcha y se ha celebrado una Jornada de Datos Abiertos para la presentación de los trabajos realizados por el grupo.

## 4.5. Recomendaciones Tecnológicas Para la Implementación de Open Data

### 4.5.1. CKAN

CKAN [4] (Comprehensive Knowledge Archive Network) es por excelencia el gestor de contenidos libres más utilizado en las iniciativas de datos abiertos, ya que garantiza la posibilidad de llegar a obtener un Open Data con la clasificación de cinco estrellas. Esta plataforma de código abierto desarrollada por la Open Knowledge Foundation permite almacenar, gestionar y publicar los datos de una organización además de facilitar la interacción con otros usuarios, personalizar o ampliar los datos. Cada conjunto de datos tiene su propia página para facilitar la búsqueda en el catálogo, proporciona una gran cantidad de características gracias a las más de doscientas extensiones. Organismos como los gobiernos de Estados Unidos, Reino Unido o la Unión Europea entre muchos otros como puede verse en la siguiente imagen utilizan esta plataforma.

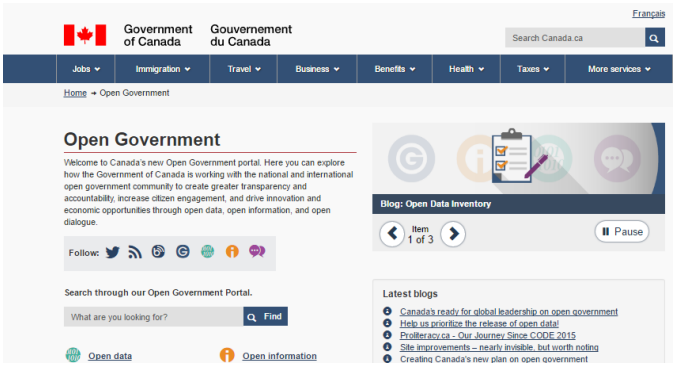


Figura 4.4: Portal de datos abiertos del gobierno de Canadá
Fuente: http://open.canada.ca/en 2017

### 4.5.2. DKAN

DKAN [5] es una herramienta que permite cargar, analizar, almacenar, publicar, catalogar y visualizar gran cantidad de datos de manera sencilla. Está basada en Drupal,y mantenida por Nucivic tiene una licencia GNU y está formado por tres componentes principales: Dkan Distro, Dkan Dataset, Dkan Datastore [14]. Algunas de sus principales características son la personalización de la apariencia de la web, facilitar la participación de usuarios, creación y gestión de metadatos propios, gestión de usuarios, visualización de datos de manera gráfica, creación

y/o importación de conjuntos de datos entre otros. La siguiente figura muestra un ejemplo de un portal desarrollado con Dkan.


Figura 4.5: Portal de datos abiertos del gobierno de California
Fuente: http://getdkan.com/img/portfolio/screencap-ca.png 2017

Estas tecnologías, ante la falta de criterio único establecido pueden considerarse estándares a la hora de desarrollar portales Open Data, por su grado de madurez y consolidación a lo largo del tiempo.

# Capítulo 5

# Conclusiones y Trabajo Futuro

## 5.1. Conclusiones

En este trabajo se han introducido los conceptos de Big Data, Cloud Computing y Open Data desde el punto de vista de la auditoría de proyectos con estas tecnologías, y la complejidad que entraña ya que en la mayoría de los casos involucra ámbitos internacionales que se ven afectados de distintas regulaciones. Se ha proporcionado una visión cercana para el usuario, explicando las medidas que se están tomando en cuanto a la seguridad de la información, y el trato que reciben los datos de carácter personal analizando la normativa europea. También se ha introducido el concepto de anonimización haciendo un breve recorrido por las principales técnicas usadas como herramienta fundamental para la seguridad de los datos y se han proporcionado formularios de ayuda para el auditor de este tipo de proyectos. Estos formularios están disponibles en: http://aws-website-tfg-4d7y9.s3-website-us-east-1.amazonaws.com/

## 5.2. Conclusions

In this paper we have introduced the concepts of Big Data, Cloud Computing and Open Data from the point of view of the audit of projects with these technologies, and the complexity involved since in most cases involves international domains that are seen affected by different regulations. A close view has been provided to the user, explaining the

measures being taken in terms of information security, and the treatment of personal data by analyzing European regulations. The concept of anonymization has also been introduced, giving a brief tour of the main techniques used as a fundamental tool for data security, and help forms have been provided to the auditor of this type of project. These forms are available at: http://aws-website-tfg-4d7y9.s3-website-us-east-1.amazonaws.com/

## 5.3. Trabajo Futuro

Como siguiente paso en este trabajo se propone realizar un seguimiento de los proyectos auditados para alcanzar la madurez de la estandarización y consolidar las conclusiones obtenidas de este trabajo.

## 5.4. Future Work

As a next step in this work it is proposed to follow the audited projects to reach the maturity of the standardization and to consolidate the conclusions obtained from this work.

# Bibliografía


[1] Abella, A. 'Reutilización de información pública y privada en España'Rooter Anal. SL, 2011.

[2] Agencia Española de Protección de Datos https://www.agpd.es, 2017

[3] http://aws-website-tfg-4d7y9.s3-website-us-east-1.amazonaws.com/

[4] https://ckan.org 2017

[5] http://getdkan.com/ 2017

[6] datos.gob.es, 2017

[7] Guías de Seguridad de Áreas Críticas en Cloud Computing, Cloud Security Alliance, www.cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf

[8] Grupo de Trabajo sobre protección de datos del articulo 29, Dictamen 05/2014 sobre técnicas de anonimización. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216$_e$*s.pdf*, 2017

[9] http://www.telegraph.co.uk/technology/2016/07/20/google-cut-its-electricity-bill-by-40pc-using-artificial-intelli/

[10] Ínformation Technology, Big Data Preliminary Report', ISO/IEC JTC1, 2014.

[11] Meloda, http://www.meloda.org/, 2017

[12] Melendrez Moreto,Ignacio, Auditoría y propuesta de metodología para publicación de datos abiertos en ciudades inteligentes, 2016.

[13] Métrica 5 Estrellas, http://5stardata.info/es/, 2017

[14] http://mauroattardi.com/dkan-una-plataforma-open-data-basada-en-drupal/ 2017

[15] Norma UNE 178301, Ciudades Inteligentes. Open Data

[16] Norma/ISO 27002, 2015.

[17] Norma/ISO 27033, 2015.

[18] http://noticias.juridicas.com/conocimiento/articulos-doctrinales/4922-iquest
existe-de-verdad-la-anonimizacion-el-grupo-del-articulo-29-de-proteccion-de-datos-no-lo-pone-facil/, 2017

[19] Open Knowledge Foundation, www.okfn.org, 2017

[20] Castaño,P. 'Big Data en la Gestión de Registros de Auditoría' http://www.eulen.com/newsletter/articulos/02$_{B}igData_{5}mayo$2015.*pdf*.

[21] Enisa, 'Privacy by Design in Big Data, an overview of privacy enhancing tecnologies in the era of big data analitycs', Enisa, December 2015s, www.enisa.europe.eu.

[22] Llamocca Portela, P., Integración y Visualización de Datos Abiertos Medioambientales, Junio 2016.

[23] 'Responsible Data Science', www.responsibledatascience.org, 2017

[24] Security Framework for Governmental clouds.

[25] Sunlight Foundation, https://sunlightfoundation.com/, 2017

# Parte I

# Apéndices

# Big Data

Cuestiones clave en proyectos Big Data

1. **Existe un perímetro de seguridad física que protege las áreas que contienen información crítica o sensible.**
   *Marca solo un óvalo.*
   - Sí
   - No

2. **El equipo está protegido ante fallos eléctricos.**
   *Marca solo un óvalo.*
   - Sí
   - No

3. **Se protegen lo cables de alimentación y telecomunicaciones contra daños ambientales y/o interferencias.**
   *Marca solo un óvalo.*
   - Si
   - No

4. **Existe y esta disponible el registro de incidencias y reparaciones de equipos.**
   *Marca solo un óvalo.*
   - Sí
   - No

5. **Existe un plan contra amenazas externas y medioambientales**
   *Marca solo un óvalo.*
   - Sí
   - No

6. **Existen métodos de eliminación de la información segura que favorecen la reutilización de los equipos.**
   *Marca solo un óvalo.*
   - Sí
   - No

7. **Se clasifica la información adecuadamente teniendo en cuenta sus requisitos legales y su nivel de sensibilidad.**
   *Marca solo un óvalo.*
   - Si
   - No

8. **El esquema de clasificación de la información es consistente para toda la organización.**
   *Marca solo un óvalo.*

   - Si
   - No

9. **La seguridad de la información está integrada en todas las fases del proyecto.**
   *Marca solo un óvalo.*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ |

10. **Los objetivos de seguridad de la información están incluidos en los objetivos del proyecto.**
    *Marca solo un óvalo.*

    - Sí
    - No

11. **La evaluación de riesgos se realiza en etapas tempranas del proyecto.**
    *Marca solo un óvalo.*

    - Sí
    - No

12. **Se analiza de forma regular las implicaciones de la seguridad de la información.**
    *Marca solo un óvalo.*

    - Sí
    - No

13. **La seguridad de la información se define y se asigna al científico de datos.**
    *Marca solo un óvalo.*

    - Si
    - No

14. **La información confidencial se borra de manera segura.**
    *Marca solo un óvalo.*

    - Sí
    - No

15. **Existen y se aplican procedimientos para la identificación de elementos que requieran un borrado seguro .**
    *Marca solo un óvalo.*

    - Sí
    - No

16. **Existe un registro del borrado de información sensible.**
*Marca solo un óvalo.*

- Sí
- No

17. **Existe una política para la protección de la información mediante técnicas criptográficas y de anonimización.**
*Marca solo un óvalo.*

- Si
- No

18. **Existen y se aplican controles contra el malware.**
*Marca solo un óvalo.*

- Sí
- No

19. **Los datos son recogidos de manera justa y legal.**
*Marca solo un óvalo.*

- Si
- No

20. **Los datos se procesan con el consentimiento previo del individuo.**
*Marca solo un óvalo.*

- Sí
- No

21. **Los datos se recogen para un fin concreto y legítimo definido de antemano.**
*Marca solo un óvalo.*

- Sí
- No

22. **Los datos procesados son los estrictamente necesarios para el propósito para el que fueron recogidos.**
*Marca solo un óvalo.*

- Si
- No

23. **Los datos de carácter personal se almacenan por un período de tiempo que no sea superior al mínimo necesario para el propósito por el cual fueron recogidos. Pasado este tiempo son eliminados de manera segura.**
*Marca solo un óvalo.*

- Si
- No

24. **El consentimiento de los individuos es completamente voluntario, habiendo sido éste previamente informado de manera específica y no ambigua.**
*Marca solo un óvalo.*

- Si
- No

25. **Se procesan DCP cuando es necesario sin interferir o entrar en conflicto con los derechos y libertades fundamentales del individuo.**
*Marca solo un óvalo.*

- Si
- No

26. **El científico de datos implementa técnicas para la protección y organización de los datos.**
*Marca solo un óvalo.*

- Si
- No

27. **El científico de datos implementa controles adecuados para limitar el acceso indebido a los datos.**
*Marca solo un óvalo.*

- Sí
- No

28. **Existe una lista con la información pertinente sobre la que el titular de los datos debe ser informado.**
*Selecciona todos los que correspondan.*

- [ ] Datos del científico de datos
- [ ] Bases legales del procesamiento y/o intereses legítimos perseguidos
- [ ] Existencia del derecho a la portabilidad
- [ ] La lista no existe

29. **Cualquier interesado puede obtener información concreta sobre la lógica que se utiliza en los tratamientos automatizados de sus datos.**
*Marca solo un óvalo.*

- Sí
- No

30. **El científico de datos toma las medidas necesarias si el usuario solicita ejercer su derecho al olvido.**
*Marca solo un óvalo.*

- Sí
- No

31. **Se cuenta con una descripción de las operaciones de procesamiento, una evaluación de riesgos de privacidad y medios para hacer frente a esos riesgos.**
*Marca solo un óvalo.*

- Sí
- No

32. **Los fines y medios del tratamiento de los DCP se determinan por dos o más científicos de datos.**
*Marca solo un óvalo.*

- Si
- No

33. **Existe un responsable que actúa como enlace con los usuarios.**
*Marca solo un óvalo.*

- Si
- No

# Cloud Computing

Cuestiones clave para proyectos Cloud Computing

1. **Existe un perímetro de seguridad física que protege las áreas que contienen información crítica o sensible.**
   *Marca solo un óvalo.*
   - Sí
   - No

2. **El equipo está protegido ante fallos eléctricos.**
   *Marca solo un óvalo.*
   - Sí
   - No

3. **Se protegen los cables de alimentación y telecomunicaciones contra daños ambientales y/o interferencias.**
   *Marca solo un óvalo.*
   - Sí
   - No

4. **Existe y está disponible un registro de incidencias y reparaciones de los equipos.**
   *Marca solo un óvalo.*
   - Sí
   - No

5. **Existe un plan de acción contra amenazas externas y medioambientales.**
   *Marca solo un óvalo.*
   - Sí
   - No

6. **Existen y se aplican métodos de eliminación segura de la información que favorece la reutilización de los equipos ayudando a evitar cuellos de botella.**
   *Marca solo un óvalo.*
   - Sí
   - No

7. **La información se clasifica adecuadamente teniendo en cuenta sus requisitos legales y su nivel de sensibilidd.**
   *Marca solo un óvalo.*
   - Si
   - No

8. **El esquema de clasificación de la información es consistente para toda la organización.**
*Marca solo un óvalo.*

- Si
- No

9. **Existe un control de acceso de los usuarios.**
*Marca solo un óvalo.*

- Sí
- No

10. **Existen y se emplean controles contra el malware en materia de detección, prevención y recuperación de la información.**
*Marca solo un óvalo.*

- Sí
- No

11. **Existen procedimientos disponibles y bien conocidos por los empleados para comunicar los eventos de seguridad.**
*Marca solo un óvalo.*

- Sí
- No

12. **Existen mecanismos de auto evaluación y mejora continua que permiten detectar y analizar errores recurrentes o de alto impacto.**
*Marca solo un óvalo.*

- Si
- No

13. **Existen directrices sobre conservación, almacenamiento y eliminación de la información.**
*Marca solo un óvalo.*

- Sí
- No

14. **Existen políticas de seguridad para los dispositivos móviles.**
*Marca solo un óvalo.*

- Sí
- No

15. **Existe un plan de acción que determina que hacer ante un cuello de botella.**
*Marca solo un óvalo.*

- Si
- No

16. **Los servicios de red se supervisan frecuentemente.**
*Marca solo un óvalo.*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ |

17. **Se usan técnicas criptográficas y de anonimización que protegen la confidencialidad, integridad y autenticidad de la información.**
*Marca solo un óvalo.*

- ◯ Si
- ◯ No

18. **Existen procesos que identifican componentes que son críticos de mantener y que por lo tanto requieren más atenciaón.**
*Marca solo un óvalo.*

- ◯ Si
- ◯ No

19. **Existe un plan de recuperación de desastres.**
*Marca solo un óvalo.*

- ◯ Sí
- ◯ No

20. **Los controles de continuidad de la seguridad se revisan regularmente.**
*Marca solo un óvalo.*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ |

# Open Data

Cuestiones clave en proyectos Open Data

1. **La información publicada es gratuita.**
   *Marca solo un óvalo.*
   - Si
   - No

2. **Existe una unidad responsable de la apertura de datos.**
   *Marca solo un óvalo.*
   - Si
   - No

3. **La unidad responsable de la apertura de datos cuenta con la formación adecuada para llevar a cabo esta tarea.**
   *Marca solo un óvalo.*
   - Si
   - No

4. **Se ofrece una API o lenguaje de consulta.**
   *Marca solo un óvalo.*
   - Si
   - No

5. **El acceso web a la información requiere registro.**
   *Marca solo un óvalo.*
   - Si
   - No

6. **No hay acceso web a la información o los datos no están en formato digital.**
   *Marca solo un óvalo.*
   - Si
   - No

7. **Reutilización.**
   *Marca solo un óvalo.*
   - Copyright
   - Uso privado
   - Reutilización no comercial
   - Sin restricciones

8. **El formato de los datos no tiene una licencia abierta.**
   *Marca solo un óvalo.*

   Si
   No

9. **Es necesario software propietario para poder extraer la información.**
   *Marca solo un óvalo.*

   Si
   No

10. **Los datos están integrados en la web.**
    *Marca solo un óvalo.*

    Si
    No

11. **Los datos están integrados y se hace referencia a otras bases de datos.**
    *Marca solo un óvalo.*

    Si
    No

12. **Existen esquemas legales técnicos que facilitan la publicación de datos.**
    *Marca solo un óvalo.*

    Si
    No

13. **Existe y se usa un control de versiones.**
    *Marca solo un óvalo.*

    Si
    No

14. **La información es procesable informáticamente.**
    *Marca solo un óvalo.*

    Si
    No

15. **Los datos se publican de manera puntual.**
    *Marca solo un óvalo.*

    Si
    No

16. **La información es primaria (cuenta con el mayor nivel de granularidad posible).**
    *Marca solo un óvalo.*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ |

17. **La información es completa.**
*Marca solo un óvalo.*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ |

18. **La información es fácilmente accesible.**
*Marca solo un óvalo.*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ |

19. **Se puede reutilizar la información libremente.**
*Marca solo un óvalo.*

- ◯ Si
- ◯ No

20. **Existe un inventario de información reutilizable elaborado por el equipo de datos abiertos y documentado.**
*Marca solo un óvalo.*

- ◯ No existe
- ◯ Existe pero no está actualizado
- ◯ Existe y esta actualizado pero no está documentado
- ◯ Existe, está actualizado y documentado

21. **Existe y se usa un método de auto evaluación y mejora continua.**
*Marca solo un óvalo.*

- ◯ Si
- ◯ No

22. **La información publicada cumple con las características del Open Data.**
*Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| No | ◯ | ◯ | ◯ | ◯ | ◯ | Si |